Towards Inherently Interpretable Machine Learning for Healthcare

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Jonas Christian Ditz
aus Berlin

Tübingen
2023

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	25.01.2024
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Nico Pfeifer
2. Berichterstatter:	Prof. Dr. Sven Nahnsen

Eidesstattliche Erklärung

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel

"Towards Inherently Interpretable Machine Learning for Healthcare"

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.5.2000) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Ort, Datum Unterschrift

Contents

Acknowledgments

To properly thank all the people involved in allowing me to successfully write this thesis would require another book in itself. Therefore, I will restrain myself to a few selected MVPs. First, my gratitude goes out to my advisor Nico Pfeifer. Throughout my entire PhD journey at the University of Tübingen, Nico stood by my side with valuable advice. I will treasure the countless discussions about my research but, especially, the ones about science in general, academia, society, and every possible other topic that came up in our meetings. While my PhD journey ends at the University of Tübingen, it started at Graz University of Technology and I would like to thank Gernot Müller-Putz for giving me my first research position in academia and all the opportunities that came with this position.

Throughout the years, I was fortunate enough to meet and collaborate with amazing researchers and human beings. I would like to thank Andreas Schwarz for his support with my research in Graz and making me feel at home in Austria. Furthermore, I would like to thank Reinmar Kobler, Joana Pereira, and all the other members at the Institute of Neural Engineering. After leaving Graz, I found a home at the Methods in Medical Informatics group and I would like to thank all members of this research group for making the last years of my professional life very special. Especially, I would like to thank Jacqueline Wistuba-Hambrecht, Florian König, Julia Hellmig, and Sofiane Ouaari. Together with Athina Gavriilidou and Anupam Gautam I was allowed to act as PhD representative at the Institute for Bioinformatics and Medical Informatics (IBMI) and would like to thank both of them for productive years and successfully carrying through many projects. For always lending a helping hand with administrative issues, I would like to express my gratitude to Agnes Molden in Tübingen and Petra Still in Graz. Additionally, I would like to thank Sven Nahnsen for agreeing to be the second reviewer for my doctoral thesis. Finally, I want to thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC number 2064/1—Project number 390727645 "Machine Learning: New Perspectives for Science" and Graz University of Technology for funding the research that resulted in this thesis.

Another huge part in finishing this doctoral thesis was played by all the friends and family members that accompanied me on this long and (sometimes) exhausting journey. First, I want to thank all my band mates at L'Apero, Thomas, Jörg, Stephanie, and Guillaume, for all the energizing and amazing rehearsals, concerts and musical moments. Second, I want to thank my parents Hans-Joachim and Monika and my siblings Lucas, Sophia, Simeon, Jakob, and Aylin for their unfaltering support during my PhD. Finally, I am for ever grateful to my girlfriend Denise for unconditionally being there in the good and, especially, the bad times.

To all of you mentioned and the many, many more that I was not able to mention by name: Thank you!

Abstract

Research in medicine and healthcare utilizes high-dimensional, multi-modal and complex data to stratify patients for improving diagnosis and treatment and discovering novel insights into diseases. Machine learning has the potential to positively impact medicine and healthcare due to the ability of machine learning methods to find complex patterns in high-dimensional data and correlate these patterns with relevant endpoints. However, substantial advancements in these two research fields can only be achieved by an interdisciplinary effort involving physicians, machine learning experts, data scientists, and more. One approach that utilizes expertise from all involved parties while also enabling scientists from different disciplines to interact on equal footing is the use of inherently interpretable prediction models. This thesis explores the viability of inherently interpretable prediction models for research in medicine and healthcare. The contributions of this thesis lie in two different fields: data-centric and model-centric research. On the data-centric side, this thesis explores the prerequisites that data have to fulfill to allow the training of inherently interpretable models. Since this type of prediction model needs to be conceptually simple, data has to fulfill quality requirements to allow the training of high performing models that are inherently interpretable. Results presented in this thesis show that a carefully designed recording setup allows to distinguish perturbation-evoked potentials from ongoing electroencephalography with simple linear models. Furthermore, a carefully curated dataset allows to train specialized prediction models for the classification of *Plasmodium falciparum*-specific protein antigen candidates. These models vastly outperform more complex prediction services for similar tasks. On the model-centric side, this thesis explores the possibility to increase the expressiveness of inherently interpretable models for prediction tasks in medicine and healthcare. The results of this thesis show that a combination of interpretable kernel functions with artificial neural networks creates inherently interpretable kernel networks that achieve state-of-the-art prediction performance on tested prediction tasks. Furthermore, these kernel networks can be robustly and efficiently trained on all dataset sizes. Since available data in medicine and healthcare ranges from small- to large-scale, this property is important for models that aim to be generally applicable in the targeted research areas. With the models proposed and the results presented in the included research, this doctoral thesis advances current knowledge towards inherently interpretable machine learning for medicine and healthcare.

Zusammenfassung

Für die Forschung im Bereich der Medizin und der Gesundheitsversorgung werden hochdimensionale, multimodale und komplexe Daten benutzt, um eine detailliertere Unterteilung von Patienten zu ermöglichen. Dies hat zum Ziel, die Diagnose und Behandlung von Patienten zu optimieren und neuartige Erkenntnisse über Krankheiten zu gewinnen. Maschinelles Lernen kann einen potenziell positiven Einfluss auf diese beiden Forschungsfelder haben, da Prädiktionsmodelle des Maschinellen Lernens die Fähigkeit haben, komplexe Muster in hochdimensionalen Daten zu erkennen und diese mit relevanten Endpunkten in Korrelation zu setzen. Um einen signifikanten Fortschritt im Bereich der Medizin und der Gesundheitsversorgung zu erreichen, ist eine interdisziplinäre Anstrengung unter Beteiligung von Ärzten, Experten des maschinellen Lernens, Datenwissenschaftlern und weiteren Interessensgruppen erforderlich. Inhärent interpretierbare Prädiktionsmodelle können ein wichtiger Ansatz für solch interdisziplinäre Forschung sein, da sie die Expertise der verschiedenen Fachleute vereint und so eine Kooperation auf Augenhöhe unterstützt. Diese Doktorarbeit untersucht die Realisierbarkeit von inhärent interpretierbaren Modellen für die Forschung in Medizin und Gesundheitsversorgung. Hierbei liegen die wissenschaftlichen Beiträge dieser Arbeit in zwei Forschungsfeldern: Datenzentrierte und modellzentrierte Forschung. Im datenzentrierten Teil werden die Voraussetzungen untersucht, die Daten erfüllen müssen, um das Training von inhärent interpretierbaren Modellen zu ermöglichen. Da diese Art von Prädiktionsmodellen eine konzeptuelle Einfachheit erfordern, muss eine besonders hohe Qualitätsanforderung an Trainingsdaten gestellt werden, um ein erfolgreiches Training mit entsprechender Prädiktionsperformanz zu erreichen. Die Ergebnisse dieser Arbeit zeigen, wie ein entsprechend ausgearbeitet Aufnahmeszenario ermöglicht, unter Verwendung von linearen Modellen sogenannte Perturbation-Evoked Potentials vom Ruhe-EEG zu unterschieden. Außerdem wird gezeigt, wie speziell kuratierte Daten das Training von Modellen für die Vorhersage von Antigenkandidaten gegen *Plasmodium falciparum* ermöglichen. Modelle, die auf diesen spezialisierten Daten trainiert wurden, erreichen eine signifikant bessere Prädiktionsperformanz als komplexere Modelle, die auf nicht kuratierten Daten trainiert wurden. Die modellzentrierte Forschung dieser Doktorarbeit untersucht die Möglichkeit, die Ausdrucksstärke von inhärent interpretierbaren Modellen für den Einsatz im medizinischen Kontext oder dem Gesundheitswesen zu verbessern. Die in dieser Arbeit vorgestellten Ergebnisse zeigen, wie aus einer Kombination aus interpretierbaren Kernfunktionen mit künstlichen neuronalen Netzen inhärent interpretierbare Modelle entstehen. Diese Modelle erreichen State of the Art Prädiktionsperformanz auf getesteten Vorhersageaufgaben. Außerdem wird gezeigt, dass diese Modelle robust und effizient auf verschiedenen Datensatzgrößen trainiert werden können. Da die Größe von Datensätzen in Medizin und Gesundheitsversorgung stark schwankt, ist diese Eigenschaft eine Voraussetzung für einen weiträumigen Einsatzbereich. Die in dieser Arbeit vorgeschlagenen Modelle und präsentierten Ergebnisse stellen einen wissenschaftlichen

Fortschritt für die Forschung zu inhärent interpretierbaren maschinellem Lernen in Medizin und Gesundheitsversorgung dar.

List of Figures

List of Tables

Acronyms

ANN Artificial Neural Network

auPR Area Under the Precision-Recall Curve

auROC Area Under the Receiver Operating Characteristic Curve

CKN Convolutional Kernel Network

CMKN Convolutional Motif Kernel Network

CNN Convolutional Neural Network

COmic Convolutional Omics Kernel Network

DMFS Distant Metastasis Free Survival

EEG Electroencephalography

MCC Matthew's Correlation Coefficient

ML Machine Learning

PAM Position-Aware Motif Kernel

PEP Perturbation-Evoked Potential

PIK Pathway-Induced Kernel

PIMKL Pathway-Induced Multiple Kernel Learning

PlasmoFAB *Plasmodium falciparum*-Specific Antigene Candidate Benchmark

RFS Relapse Free Survival

RKHS Reproducing Kernel Hilbert Space

RL Reinforcement Learning

SL Supervised Learning

SLDA Shrinkage Linear Discriminant Analysis

SSL Self-Supervised Learning

XAI Explainable Artificial Intelligence

1 List of Publications

This cumulative doctoral thesis is based on the manuscripts listed below. The order of the listed publications is determined by topic rather than the chronological order. Since each work was a collaboration of several scientists, the following pages are dedicated to indicate my personal contributions.

Accepted Publications

1. **Ditz, Jonas C.**, Schwarz, Andreas, and Müller-Putz, Gernot R. "Perturbation-evoked potentials can be classified from single-trial EEG." *Journal of neural engineering* 17.3 (2020): 036008.
2. **Ditz, Jonas C.***, Wistuba-Hamprecht, Jacqueline*, Maier, Timo, Fendel, Rolf, Pfeifer, Nico, and Reuter, Bernhard. "PlasmoFAB: a benchmark to foster machine learning for Plasmodium falciparum protein antigen candidate prediction." *Bioinformatics* 39.Supplement_1 (2023): i86-i93.
3. **Ditz, Jonas C.**, Reuter, Bernhard, and Pfeifer, Nico. "Inherently interpretable position-aware convolutional motif kernel networks for biological sequencing data." *Scientific Reports* 13, 17216 (2023)
4. **Ditz, Jonas C.**, Reuter, Bernhard, and Pfeifer, Nico. "COmic: convolutional kernel networks for interpretable end-to-end learning on (multi-) omics data." *Bioinformatics* 39.Supplement_1 (2023): i76-i85.

* indicates equal contribution

2 Introduction

Why are you using models that you do not understand to investigate something that you do not understand? This question was asked as a response to a talk at a machine learning (ML) summer school that I attended during my doctoral studies. There were several ML experts in the audience that day. However, it took one of the few non-experts to ask this question. In retrospective, the answer that was provided by the present ML experts was even more striking: the world is complex and, hence, we have to increase the complexity of models as much as possible. This answer has to be noteworthy for scientists as it seems to be in stark contrast to the law of parsimony, more famously known as Occam's razor*, a philosophical concept that has supported the falsifiability criterion of the scientific method for centuries [1]. Yet there are large parts of the machine learning community that work towards increasing complexity, just as suggested by the answer [2, 3, 4]. While these models' achievements that were rightfully celebrated seem to validate this approach of increased complexity, simpler but understandable models offer a key benefit to both researching scientists and users in an application scenario: they can help them to understand the problem at hand. Since gaining an understanding about the investigated research subject is fundamental in healthcare, one of the key questions I investigated during my doctoral studies was what the prerequisites for understandable models in healthcare are.

In the digital world, living life is equivalent to generating data. Internet traffic is closely monitored and used to automatically generate profiles about users. Most governments create and store information about their citizens ranging from occupation and place of residency to health status. The progress of pupils and students in schools and universities is closely tracked and analyzed. Even our body consist of and constantly generates data that we are able to extract and store. The genomic information can be accessed using sequencing techniques and several other methods exist to access additional information, like the transcriptome, proteome, metabolome, microbiome, or phenome. Almost everything in the digital world either produces and/or stores data. While this data can be seen as a discretization and simplification of reality, it still is usually high-dimensional and, hence, seemingly of high complexity. Since humans cannot comprehend high-dimensional data, considerable effort is put into extracting humanly comprehensible information from data. In the field of machine learning, this effort can be roughly separated into three categories. Unsupervised or self-supervised learning (SSL) describes methods that try to group together samples based solely on characteristics of the

William of Ockham (1287, †10 April 1347) was an English philosopher and theologian

observed data. The main idea is that the learned groups and characteristics, often referred to as patterns, are a mimicry of reality and therefore act as a comprehensible representation of the real world. On the other hand, supervised learning (SL) describes methods that correlate patterns with comprehensible output variables, often referred to as labels. These labels can be continuous for regression problems or categorical for classification problems. The third category, reinforcement learning (RL), summarizes methods that try to learn a sequence of actions based on complex data. While self-supervised and reinforcement learning were mentioned for the sake of completeness, the scope of the remaining thesis will focus on supervised learning.

The desire to create increasingly complex models can be intuitively understood by looking at the concept that embodies the core of ML theory, called empirical risk minimization (ERM) [5]. In plain English, ERM is used to minimize the error produced by a model for a specific dataset. To introduce this concept more formally, assume a joint probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ that gives rise to pairs over an arbitrary domain set $\mathcal{X}$, which will be called *instance space*, and a set of labels $\mathcal{Y}$. The domain points are often called *instances*. The task for a model is to produce a prediction function $h : \mathcal{X} \to \mathcal{Y}$. The error of this prediction function is most often defined as the probability to draw a random pair $(x, y) \sim \mathcal{D}$ such that the prediction function yields a result that is different from y [6]:

$$L_{\mathcal{D}}(h) \overset{\text{def}}{=} \mathbb{P}_{(x,y)\sim\mathcal{D}}\left[h(x) \neq y\right] \overset{\text{def}}{=} \mathcal{D}\left(\{(x, y) : h(x) \neq y\}\right). \tag{2.1}$$

Since the distribution $\mathcal{D}$ is unknown to the model, ERM utilizes a sequence of known pairs of instances with corresponding labels[†] $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, $(x_i, y_i) \sim \mathcal{D}$ to estimate the model's error [6]:

$$L_{\text{ERM}}(h) \overset{\text{def}}{=} \frac{|\{i : h(x_i) \neq y_i\}|}{m}, \tag{2.2}$$

where $i \in \{1, \ldots, m\}$. The obvious goal of utilizing machine learning is to find a labeling function that is as close as possible to the distribution that gives rise to the labeled instances. Given the notion established in the previous paragraph that the data generating process of the world is highly complex, it only seems consequential to find an equally complex prediction function[‡]. Furthermore, the ERM paradigm together with basic statistics resulted in a plethora of different performance metrics that can be used by researchers to quantify the validity of their complex models ranging from simple accuracy to more complex measures like the area under the receiver operating characteristic curve (auROC) [8] or Matthew's correlation coefficient (MCC) [9].

From a scientific point of view, there are three main issues that arise from increasing model complexity to improve prediction performances. The first more general and philosophical problem comes from the modern machine learning practice to use performance metrics not only for evaluating models but also as a target during training. This procedure results in a questionable validity of performance metrics as an evaluation tool for ML models, an idea that

[†]This sequence of domain points is often called a *training set*. However, these sequences are not guaranteed to fulfill the requirements of a set, e.g., some instances may occur multiple times and the order of instances can be taken into account by some algorithms. [6]

[‡]In the classical machine learning literature, the bias-complexity trade-off is usually used to advise against overly complex models as it can lead to overfitting on training data [6]. However, empirical results suggest that this limitation is not prevalent for one type of models that is at the forefront of highly complex and over-parametrized learners: (deep) artificial neural networks (ANNs) [7].

is famously verbalized in the adage Goodhart's law[§]. The second issue has a more practical nature. The world in its entirety might be highly complex but (supervised) learning tasks are not performed on the entire world. Usually only a tiny part of all available information is needed for a specific prediction task. And while this information can still be high-dimensional (e.g., gene expression data), we often have prior knowledge about the data that reduces the complexity (e.g., gene interaction networks for gene expression data). The singular focus on increasing model complexity to compensate for high-dimensional data disregards previously attained scientific knowledge and raises the danger of learning spurious correlations [12, 13]. And the last issue comes from the fact that using a highly complex, black-box model to compute predictions from data points renders it impossible to understand the rationale behind the assignment of predicted outcomes directly from the model. This prevents one of the most crucial aspects of conducting science: knowledge discovery. Using prior knowledge to create intrinsically interpretable models offers a solution to all three of these issues by (i) providing an additional mode for model evaluation, (ii) simpl++ifying the input data domain with previously attained knowledge, and (iii) opening the black box to provide means for understanding the rationale behind a prediction.

2.1 Preliminaries

Throughout this thesis, different concepts will be mentioned. This section serves as a summary for the collection and introduction of important concepts.

2.1.1 Performance Metrics

As already mentioned, performance metrics serve as an indicator of a prediction model's success. Although there are numerous metrics introduced in ML literature, only metrics that are relevant for the remaining thesis will be described here. Furthermore, the binary version of these metrics will be introduced. All of the following performance metrics can be derived using four basic quantities. True positives (TP) describe the number of data points that belong to the positive class and are correctly classified by the model. False positives (FP) are data points that belong to the negative class but are wrongly classified by the model. True negatives (TN) are all data points belonging to the negative class that are correctly classified by the model. Finally, false negatives (FN) describe the number of data points that belong to the positive class but are wrongly classified by the model.

Accuracy is the most basic performance metric. It indicates the number of correctly classified samples and is defined as [14]

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \tag{2.3}$$

A slightly more advanced version is called **balanced accuracy**. This metric tends to perform better on imbalanced data and has different definitions in the literature. For this thesis, the

[§]Goodhart's law states: "When a measure becomes a target, it ceases to be a good measure". [10, 11]

balanced accuracy is defined as the arithmetic mean of sensitivity and specificity [15]:

$$\mathrm{Acc_{bal}} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right). \tag{2.4}$$

Another commonly used metric is the $\mathbf{F_1}$ **score**, which is defined as the harmonic mean of precision and recall [16]:

$$\mathrm{F_1} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}. \tag{2.5}$$

Matthew's correlation coefficient (MCC) is a metric that puts equal emphasis on all four of the basic quantities and provides a more robust measure than the previous mentioned metrics. It is also widely accepted as one of the most reliable performance metrics for biological data and defined as [9]

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN}-\mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}. \tag{2.6}$$

Additionally to these mathematically motivated metrics, there are graphically motivated performance metrics. Probably the most widely used metric in ML literature is called the **area under the receiver operating characteristic curve** (auROC) [8]. The ROC curve describes a plot that shows the change of true positive rate ($\mathrm{TNP} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$) against false positive rate ($\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$) for different discrimination thresholds of a model [17]. In order to compare different models, this curve is often reduced to a single number that describes the area under the ROC curve. Although there are different methods to discretize the auROC calculation, the well-known utilization of a *Wilcoxon-Mann-Whitney* statistic [18] was used in this thesis:

$$\mathrm{auROC} = \frac{\sum_{x^-\in N^-}\sum_{x^+\in N^+}\mathbf{1}\left[h(x^-) < h(x^+)\right]}{(\mathrm{TN}+\mathrm{FP})\,(\mathrm{TP}+\mathrm{FN})}, \tag{2.7}$$

where N^- denotes the set of negative samples and N^+ denotes the set of positive samples, and $\mathbf{1}\left[h(x^-) < h(x^+)\right]$ is an indicator function that returns 1, if the prediction for the negative sample is smaller than the prediction for the positive sample and 0 otherwise. Since it becomes more widely known that the auROC estimate is overly optimistic in many scenarios [19], it becomes increasingly common to use a different graphically motivated metric called the **area under the precision-recall curve** (auPR). The PR curve plots precision ($\mathrm{Pr} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$) against recall ($\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$) at different discrimination thresholds of a model. However, due to the properties of the PR space, the auPR cannot be estimated using linear interpolation methods [20]. All auPR values in this thesis are approximated with the average precision (AveP), which is a commonly used, robust approximation for the auPR [21]:

$$\mathrm{auPR} \approx \mathrm{AveP} = \sum_{t\in T}\Delta\mathrm{R}_t\,\mathrm{Pr}_t, \tag{2.8}$$

where T is the set of all thresholds, Pr_t is the precision at the t-th threshold, and $\Delta\mathrm{R}_t$ is the change in recall between the t-th and the previous threshold.

2.2 Let's Talk About Data

At the core of every single machine learning application lies data. One can develop the most sophisticated, intelligent new method, data will be used to judge the usability of this method. Even the most advanced machine learning model cannot compensate for a lack of informative structure within a dataset. And attempts to interpret a model are futile if the data cannot be interpreted due to poor quality. Nevertheless, the quality of data was mostly ignored by the machine learning community in the past, which is clearly noticeable from the fact that it took until 2021 before one of the most important machine learning conferences, the Conference on Neural Information Processing Systems (NeurIPS), introduced a track that focuses on data. There are still highly active sub-fields of machine learning research that consider the quantity of data more important than the quality, e.g., reflected by the common practice of the natural language processing (NLP) community to scrap huge amounts of unfiltered text data from the internet [22] or the practice to exploit underpaid workers to create large, labeled image datasets with minimal quality control [23].

While the quality of data might be a minor concern in economical applications of machine learning, where the main incentive is the maximization of monetary interests, the usage of high-quality data is non-negotiable in high-stakes scenarios like healthcare. Two aspects can be identified that play an important role in determining the quality of data for healthcare-related ML applications. For this thesis the two aspects will be called *specificity* and *completeness*. The first aspect, *specificity*, points at the fact that a dataset has to contain information that is specific for the targeted prediction task. A general limitation of ML models is the inability to guarantee an acceptable prediction performance for data that is generated from a different distribution than the training data. This limitation is called a lack of out-of-distribution generalization [24]. Furthermore, a model's prediction performance tends to decrease if that model is trained on datasets which contain a lot of unrelated information or "noise" [25]. However, since the process of data generation is laborious and expensive, especially for healthcare-related data, it is a common practice to either reuse pretrained models or reuse data for different purposes [26, 27, 28, 29, 30, 31, 32]. While there should not be a general advise against these practices, e.g., they can help to significantly decrease resource usage, great care has to be taken when reusing models or data. The process often relies on complex mathematical modeling that decreases the understandability of resulting models [33]. The second aspect, *completeness*, describes whether a dataset contains enough information to paint a complete picture of the targeted prediction task. For healthcare-related data, *completeness* does not exclusively mean that data contains the targeted information, e.g., samples from the gene that is supposed to be investigated or to ensure that every important protein is included in the dataset. The balanced representation of meta-information, like ethnicity, gender, or age, is equally important. Excluding minorities can lead to an incomplete knowledge about the researched subject, e.g., diseases or treatment results, and biased prediction models [34, 35, 36, 37].

However, similar to the fact that modern science is rarely advanced by individuals but in a community effort, relying on an individual or a single research group for the judgment of a data's specificity and completeness is an unnecessary limitation. Utilizing the collective expertise provided by the scientific community is a much more robust approach. There are several examples showing the benefits of a community effort to judge the quality of data ranging from recidivism prediction [35, 38] over facial recognition [37] to medical imagery [36]. In order

to provide researchers with the means necessary for such an assessment of data, two main prerequisites have to be fulfilled. Researchers have to be able to find the data and they have to be able to understand the purpose of the data. In the past years, there were several efforts to establish standards for these two prerequisites. Most noticeable was the introduction of the FAIR principles that were introduced to improve the **f**indability, **a**ccessibility, **i**nteroperability, and **r**euse of data [39]. Together with the easy-to-follow FAIRification framework, researchers are provided with a simple opportunity to ensure the first prerequisite. To fulfill the second prerequisite, Gebru and colleagues developed a datasheet for datasets [40]. These datasheets provide researchers with all information necessary to understand the dataset such as details about the creation process, data characteristics, recommended uses, and other information.

2.3 A Kernel Of Truth

One of the core principles that were utilized in the development of the interpretable models introduced with this thesis is the use of kernels. In general, kernels describe a type of similarity measure and the underlying idea of the use of kernels is quite simple. Most traditional machine learning methods like support vector machines (SVMs) are using halfspaces to separate samples into classes, i.e., assuming w.l.o.g. $\mathbb{R}^n$ as the domain set $\mathcal{X}$, the prediction function can be written as $h(x) = \langle w, x \rangle + b$ with $w, x \in \mathbb{R}^n$ and $b \in \mathbb{R}$. However, realistic data is rarely separable by halfspaces due to their restricted expressive power. To overcome this limitation, all instances can be mapped into an often higher-dimensional feature space $\mathcal{F}$ by defining a non-linear mapping $\varphi : \mathcal{X} \to \mathcal{F}$. This feature space can be any *Hilbert space*[¶] including infinite-dimensional spaces. Learning a separating halfspace in the feature space results in a prediction model that is linear in $\mathcal{F}$ but non-linear in the original instance space $\mathcal{X}$. The validity of this mapping approach can be easily shown since, for every probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, it is possible to define an image probability distribution $\mathcal{D}^\varphi$ over $\mathcal{F} \times \mathcal{Y}$ using the following definition: $\forall A \subseteq \mathcal{F}, \mathcal{D}^\varphi(A) = \mathcal{D}(\varphi^{-1}(A))$. Therefore, the following equation holds for every prediction function h over the feature space: $L_{\mathcal{D}^\varphi}(h) = L_{\mathcal{D}}(h \circ \varphi)$, where $h \circ \varphi$ is the composition of the prediction function onto the mapping. [6]

Increasing the dimensionality of instances to improve the expressiveness of a model comes with an important disadvantage: the curse of dimensionality [41]. In machine learning, this term refers to two issues that arise when the dimensionality of instances becomes larger. First, under the assumption of uncorrelated dimensions, the number of instances needed to robustly train a prediction model increases with the dimensionality of instances due to the higher VC-dimension[‖] of the model. And second, the complexity of calculations in higher dimensions can result in models that are computationally infeasible to train. While there is a vast amount of literature about the first problem (and a detailed introduction of the issue would be out of scope of this thesis), the computational complexity issue can be resolved using the so-called *kernel trick*. This approach utilizes the fact that the feature space $\mathcal{F}$ has an inner product and

[¶]The term Hilbert space describes any vector space for which a distance function in form of an inner product is defined. Furthermore, the space has to be complete with regard to the distance function.

[‖]The Vapnik–Chervonenkis (VC) dimension of a prediction model is the maximum number of points that the model can shatter [42].

defines a kernel function such that, for two instances $x_1, x_2 \in \mathcal{X}$,

$$K(x_1, x_2) = \langle \varphi(x_1), \varphi(x_2) \rangle \tag{2.9}$$

is the similarity of the two instances within the feature space. Defining such kernel functions allows for learning linear prediction models in high-dimensional spaces without the need to either explicitly specify the mapping φ or use points in the high-dimensional space. The theoretical basis for this property is the well-known *representer theorem* [43], which states that a linear combination of a kernel function evaluated on a set of instances can be used to express any function that minimizes an ERM functional in the corresponding feature space. The representer theorem provides the mathematical reason why the feature space of a kernel method is usually called a *reproducing kernel Hilbert space* (RKHS)**. The findings of the representer theorem can be used to show that the only information needed to train a kernel method is the matrix $G \in \mathbb{R}^{m \times m}$ s.t., $G_{i,j} = K(x_i, x_j)$ defined over the training set. This matrix is usually called the kernel's *Gram* matrix. The use of kernel functions is possible for any prediction model that only relies on inner products, which sounds limiting but includes almost all commonly used prediction models [44, 43, 45]. Furthermore, kernel functions can be used to incorporate prior knowledge into prediction models. This is especially useful for healthcare-related data. There are established kernel functions for biological sequences, like the spectrum kernel [46, 47], the mismatch kernel [48], and the weighted degree kernel (with shifts) [49], as well as kernel functions for two- and three-dimensional molecular structures [50, 51, 52].

2.3.1 Interpretable Kernel Functions for Healthcare-Related Data

While the use of kernels can improve prediction performance by introducing non-linearity and prior knowledge to machine learning models, they are not inherently designed to make a model interpretable. If the goal is to achieve a method that provides interpretability, prior knowledge has to be carefully incorporated into the mapping and similarity measure. One kernel function that offers performance improvement and interpretability and acts as an important prior work for this thesis is the **oligo kernel** [53]. Meinicke and colleagues defined so-called oligo functions that encode the occurrence of a k-mer using a tuneable degree of positional uncertainty. The term k-mer describes strings of length k over an alphabet. The prior knowledge used by the oligo kernel is the fact that the occurrence of k-mers contain crucial information for computing the similarity of biological sequences. Furthermore, this occurrence can have a certain degree of positional uncertainty in real-world biological sequences, i.e., two sequences can be highly similar even if the occurrence of certain k-mers slightly differ between these two sequences. The authors utilized oligo functions to define the oligo kernel as

$$K_{\text{oligo}}(x_i, x_j) = \sqrt{\pi}\sigma \sum_{\omega \in A^k} \sum_{p \in S^i_\omega} \sum_{q \in S^j_\omega} \exp\left(-\frac{1}{4\sigma^2}(p-q)^2\right), \tag{2.10}$$

where x_i and x_j are biological sequences, A^k is the set of all k-mers over an alphabet A, S^i_ω is the set of starting positions of k-mer ω in sequence x_i, and S^j_ω is defined equivalently for

**This notion is not entirely precise since the representer theorem is based on the fact that the feature space of a kernel is an RKHS. However, the details are out of scope of this thesis and the interested reader is referred to the literature about the representer theorem and reproducing kernel Hilbert spaces in general.

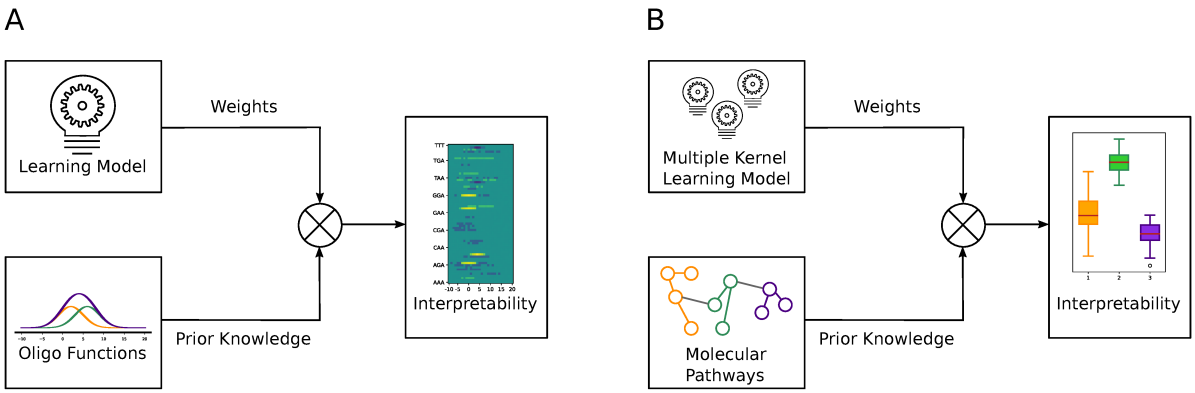

Figure 2.1: Schematic on how selected kernel methods incorporate prior knowledge to achieve interpretability. **A**: The oligo kernel encodes positional uncertainty of k-mer occurrence with so-called oligo functions. Once a prediction model, e.g., a support vector machine, is trained, the weights of the model can be used to compute the influence of each k-mer on the model's decision. **B**: The pathway-induced kernel utilizes molecular interaction networks to generate pathway-induced kernel functions. The weights of a multiple kernel learning model trained with the pathway-induced kernel functions can be used to compute the influence of each pathway on the model's decision.

sequence x_j. σ is the tuneable degree of positional uncertainty. Meinicke and colleagues showed that a linear combination of oligo functions with weights learned by a learning model can be used to compute an informative visualization that provides the means to interpret a prediction decision made by the model (see Figure 2.1 A for an exemplary visualization).

The second interpretable kernel function that influenced the work presented in this thesis is the **pathway-induced kernel** (PIK). This kernel was developed for training kernel methods on molecular measures, e.g., gene expression data. The prior knowledge utilized in the PIK comes from molecular interaction networks. Connectivity information provided by these networks is used to create a graph representation of molecular measures and the corresponding graph Laplacian [54] is used to encode similarity of different molecular measures as

$$K_{\text{PIK}}(x_i, x_j) = x_i^T L x_j, \tag{2.11}$$

where x_i and $x_j \in \mathbb{R}^n$ are molecular measures and $L \in \mathbb{R}^{n \times n}$ is the graph Laplacian of the molecular interaction network. In order to provide interpretability, Manica and colleagues divided the interaction network into pathway-specific sub networks [55]. That resulted in the pathway-induced multiple kernel learning (PIMKL) approach given as

$$K_{\text{PIMKL}}(x_i, x_j) = \{x_{i,p}^T L_p x_{j,p}\}_{p \in P} \tag{2.12}$$

where P is the set of used pathway-specific sub-networks, $x_{i,p}$ and $x_{j,p} \in \mathbb{R}^d$ are the d-dimensional slices of the molecular measures that correspond to pathway p, and $L_p \in \mathbb{R}^{d \times d}$ is the graph Laplacian of pathway p. Training a multiple kernel model using this set of kernel functions learns a weight for each pathway and a visualization that can be used to interpret a prediction decision made by the model can be computed with these weights (an exemplary visualization is shown in Figure 2.1 B).

2.4 The Good, the Bad, and the Kernel Network

Although kernel methods can handle high-dimensional data, as described in the previous section, there is an issue arising from the use of traditional kernel methods that limits their applicability to certain data. As mentioned before, the prerequisite to train a kernel method is the Gram matrix. However, calculating this matrix scales quadratically with the number of instances in the training set resulting in the inapplicability of kernel methods to datasets with a vast number of instances, which is often referred to as "big data". Although small to medium sized datasets are still more commonly found in healthcare-related fields, "big data" starts to become a crucial part of research and application due to technical advances in data-generating processes resulting in decreased prices for collecting biological data and the introduction of electronic information management systems for patient data like the electronic health record (EHR) [56, 57]. In other words, prediction models for healthcare should be ideally applicable to all sizes of datasets. Utilizing large amounts of available instances requires models to have an efficient training procedure and one type of models that are exceptionally good at handling large amounts of data are artificial neural networks (ANN) [58]. Two main factors are contributing to the scalability of ANNs. First, the training process utilizes the gradient of a loss surface and stochastic gradient methods can be used to iteratively use smaller chunks of the available instances to approximate the real gradient [58]. Second, the computations involved in training ANNs can be massively parallelized. Furthermore, there are several different types of networks available. These so-called architectures, which include but are not limited to convolutional, recurrent, or transformer networks, allow the application of ANNs on different input data resulting in a highly flexible model type [58]. Neural networks are usually trained in an end-to-end scheme that combines feature embedding and prediction in one model. The benefits of this training procedure come in form of the reduced labor that is necessary to perform additional pre-processing like feature embeddings and a harmonization of the whole prediction pipeline [59]. However, neural networks are usually deployed as heavily over-parameterized deep learners, On the one hand, this requires additional effort to robustly train such a complex model on small or medium sized datasets. Although this limits their usability on many healthcare-related prediction problems, there is a significantly increased number of published works utilizing these model types for healthcare-related tasks and indicating their potential in recent years [60, 3, 61, 62, 63]. On the other hand, there is the issue that over-parametrization results in incomprehensibly complex models that do not provide any means to directly understand the decision making process. In other words, deep models are black boxes.

One recent strain of research that aims to keep the scalability and end-to-end training capabilities of neural networks but improve their robustness on smaller datasets is the combination of neural networks and kernel methods [64, 65, 66, 67, 68, 69]. A seminal work towards kernel networks was published by Mairal and colleagues [65, 66]. One of the core ideas is the use of the Nyström method that allows to project points onto a finite-dimensional subspace of an RKHS belonging to some kernel K [70]. The subspace is defined by selecting a set of anchor points $Z = \{z_1, \ldots, z_k\}$, $z_i \in \mathcal{X}$ and define a subspace $\mathcal{E} \leq \mathcal{F}$ that is spanned by the anchor points, i.e. $\mathcal{E} \stackrel{\text{def}}{=} \text{Span}\left(\varphi(z_1), \ldots, \varphi(z_k)\right)$. Mairal used findings from [70] and [71] to introduce an explicit formula for the orthogonal projection of instances onto the subspace $\mathcal{E}$. Assuming

that all anchor points have unit l_2 norm, Mairal's projection $\psi : \mathcal{X} \to \mathcal{E}$ is defined as:

$$\psi(x) = \|x\| \; G_{ZZ}^{-\frac{1}{2}} \, G_Z \left(\frac{x}{\|x\|} \right), \tag{2.13}$$

where $x \in \mathcal{X}$ is an instance, $G_{ZZ} = (K(z_i, z_j))_{i=1,\ldots,k;j=1,\ldots,k}$ is the Gram matrix over the set of anchor points, $G_{ZZ}^{-\frac{1}{2}}$ is the (pseudo-)inverse square root of the Gram matrix, and $G_Z \left(\frac{x}{\|x\|} \right) = \left(K \left(z_1, \frac{x}{\|x\|} \right), \ldots, K \left(z_k, \frac{x}{\|x\|} \right) \right)^T$ is a vector containing the evaluation of the kernel function between the l_2-normalized instance x and every anchor point. Mairal and colleagues showed in their work that such an explicit parametrization can be incorporated into a convolutional layer that enables a neural network to learn feature representations within the subspace of an RKHS. Furthermore, this allows to define a gradient on the anchor points, which enables the tuning of anchor points within the same end-to-end learning scheme used for model training. The result is a neural network that uses the RKHS of a kernel function instead of the domain space $\mathcal{X}$ to solve a prediction task. Due to the similarity to kernel methods, these network are called kernel networks. Mairal and colleagues showed that the use of the kernel embedding allows kernel networks to be robustly trained on smaller datasets. In a following work, Chen and colleagues showed the feasibility of these kernel networks for different biological data modalities [67, 68, 69]. While these previous publications introduce a type of learning model that can be robustly applied to different sized datasets in healthcare-related fields, they focused on utilizing kernel functions that do not provide inherent interpretability. Consequentially, the resulting models required post-hoc modeling to compute interpretations.

2.5 Towards Interpretable Machine Learning for Healthcare

Interpretability is a concept that gains increasing traction in machine learning. Although it is not mathematically definable, interpretability is usually used in the machine learning community to summarize methods and algorithms that allow humans to understand the cause of a decision [72, 73, 74]. In many cases this involves analyzing the influence of specific features, e.g., pixels/superpixels in imagery or gene products in gene expression data. Another approach is to analyze model internals like weights in linear models, tree structure in decision trees, or feature detectors in ANNs. Interpretability in machine learning can be categorized by three main characteristics: (i) local vs. global, (ii) inherent/intrinsic vs. post-hoc, and (iii) model-specific vs. model-agnostic. The first characteristic describes the type of interpretation that is provided. Local interpretability allows humans to understand a model's decision made for one specific instance. Global interpretability allows humans to understand a model's behavior in general by providing an understandable decision surface. The second characteristic describes whether prediction and interpretability come from the same model. Inherently interpretable models are prediction models that allow to directly assess the decision surface and model internals to enable understanding of the model's behavior and specific decisions. These models are sometimes referred to as white boxes. Post-hoc models cannot be used to make predictions but are designed to provide the means for understanding prediction models that are incomprehensible to humans due to their complexity. As mentioned before, these types of prediction models are also known as black boxes. The third characteristic describes whether an interpretability method can be used for any type of prediction model or can only be used for specific model types.

Currently, a significant research effort in terms of interpretability is directed towards post-hoc models [75, 76, 77, 78, 79, 80, 81, 82]. The rationale behind this research effort is quite obvious: using post-hoc models to provide interpretability does not impose any restrictions on the complexity of the prediction model. In other words, the current trend of steadily increasing model complexity is not hindered by the use of post-hoc interpretability. However, there are limitations to post-hoc models that render their applicability in high-stakes scenarios, like healthcare, questionable. One issue arises from the fact that post-hoc models are unfaithful with regard to the computations done by the prediction model [83]. It is possible to show that many post-hoc methods that are utilizing the information flow through a deep model ignore deeper layers when they calculate the interpretation [84]. Due to this unfaithfulness, an interpretation that was calculated with a post-hoc model cannot guarantee to accurately reflect the cause of the decision made by a prediction model. Especially in healthcare, decisions that are made based on an unfaithful or wrong model interpretation can cause serious harm. Beside these technical issues, a general shortcoming of post-hoc interpretation is the fact that computed interpretations for a single decision can be significantly different dependent on parameter-choice and algorithm-choice due to the fact that providing post-hoc interpretability is underdetermined [85]. Currently, there is no accepted method to compare and rank the validity of two different post-hoc interpretations. While this might pose a negligible problem in research due to the commonly assumed cooperative scenario, the situation changes drastically if there are parties involved that have adversarial motivations [85]. For example, a profit-driven company offering a treatment-decision support system might have different incentives than a physician who is using the system to treat patients. As the provider of interpretability, the company can choose a method that maximizes their incentive but that does not guarantee that this method also maximizes the physician's incentive. Inherently interpretable models do not suffer from many of these limitations. They are faithful with regards to the computation, since the same model computes the prediction and interpretation. Furthermore, inherently interpretable models are using internals and the direct access to the decision surface when computing interpretations. Therefore the provided information used to understand the cause of a decision is unique. This mitigates the variability issue arising from the use of post-hoc methods. In other words, inherently interpretable models do not share the properties that make post-hoc methods questionable in high-stake scenarios[††].

This thesis advances the field of interpretable machine learning for healthcare by first showing how curated, high-quality data with validated prior knowledge can be used to reduce the necessity for complex models, one of the main prerequisites for inherently interpretable machine learning models. The focus will lie on specificity, the first aspect of data quality as defined in section 2.2. In the following chapters, the importance of data that contain validated information for a targeted prediction task will be shown. Furthermore, the results in this thesis indicate that a careful curation process allows to reuse healthcare-related data by ensuring specificity for the new prediction task. For datasets published in research that contributed to this thesis, compliance with the FAIR principles [39] is ensured and datasheets [40] are provided. Second, this thesis shows how prior knowledge can be incorporated into machine learning models to not

[††]One thing that has to be clearly stated is that none of the mentioned limitations of post-hoc interpretation should be read as a general advise against the use of black-boxes in high-stake scenarios. There are ongoing discussions about this question between scientists, politicians, companies, and interest groups and the complexity of this matter cannot be adequately addressed in this thesis.

only reduce the complexity of these models but also make them intrinsically interpretable. To this end, Mairal's work on kernel networks is utilized to combine interpretable kernel functions with ANNs to create conceptually simple models that enable users to compute global and local interpretation without the need for post-hoc methods. The presented models are applicable to two of the most commonly used data modalities in healthcare: biological sequences and tabular omics data with an underlying graphical structure.

3 Objectives

The overarching goal of my thesis was to explore the possibility to utilize inherently interpretable machine learning models for healthcare-related prediction problems that provide the means to unambiguously and faithfully interpret the cause for a decision by simultaneously achieving a prediction performance that is comparable to state-of-the-art methods. I approached this goal from the two main directions for inherently interpretable machine learning: data-centric and model-centric research.

In the data-centric part of my thesis, I investigate two main research questions. The first came from the field of brain-computer interfaces (BCIs), which is a subfield of neural engineering. I investigated perturbation-evoked potentials (PEPs). This neural activity pattern gets elicited when humans lose their balance control. Being able to robustly capture the occurrence of a PEP can provide benefits in many different applications including gait rehabilitation [86], virtual reality [87], and aviation/driving assistance systems [88]. In order to investigate the feasibility of integrating PEP-detection into systems, I developed an experimental study that allows the recording of high-quality PEP data using electroencephalography (EEG). Furthermore, I investigated the possibility to detect PEPs from high-quality data with a linear classifier for various recording setups, i.e., different electrode layouts. The second research question arose from the continuous need to develop effective drugs and vaccines to fight malaria. Although this disease caused by the parasite *Plasmodium falciparum* (Pf) is one of the most relevant infectious diseases, the utilization of computational methods to efficiently explore protein antigen candidates for drug/vaccine development is hindered by the current sparsity of Pf-specific proteins with known functionality [89, 90, 91, 92]. To address the challenge for utilizing computational methods for protein prescreening, I aimed to develop a manually curated benchmark with validated labels for the training of prediction models. Furthermore, I investigated the implications of using pre-trained prediction models.

The model-centric part of my thesis is focused on kernel networks and the possibility to make them inherently interpretable. As a first research question, I investigate whether a carefully designed kernel function can be used to develop an inherently interpretable kernel network for biological sequences. The need for a model that is specifically designed for this data type arises from the fact that biological sequences contain valuable information for various healthcare-related prediction tasks ranging from drug resistance [93] to binding affinity [60, 94]. However, in recent years it became increasingly obvious that focusing on a single biological

data type is detrimental to capturing the real structure behind many prediction tasks [95]. To provide the possibility to utilize this so-called multi-omics approach, I aimed to develop inherently interpretable kernel networks for tabular omics data. Since a majority of omics modalities are provided as tabular data, e.g., gene expression, DNA methylation, copy number variation, immunoassay, etc., providing a kernel network that can be applied on such data enables researchers to combine commonly used modalities for multi-omics prediction.

4 Results and Discussion

In the following chapter, I will summarize my research towards inherently interpretable machine learning for healthcare. First, the main ideas and results of each manuscript will be briefly presented and individually discussed. Section 4.1 focuses on the data-centric part of my thesis by presenting and discussing my research on perturbation-evoked potentials as well as *Plasmodium falciparum*-specific proteins that are potential antigens. My model-centric research is presented in section 4.2. Here I will present and discuss my research on the creation of inherently interpretable kernel networks for different biological data modalities. Both sections will relate the work I have done in the respective fields to each other and to the main goal of inherently interpretable machine learning for healthcare. An integrated discussion of my research that connects the different directions I explored and provides a joint perspective on my research towards inherently interpretable machine learning for healthcare will be presented in section 4.3.

During this chapter, I will switch between the subject pronouns *we* and *I*. If the general work published in manuscripts is presented, I will use the plural pronoun since each manuscript has several authors. If my own contributions and views are presented, I will indicate that by using the singular pronoun.

4.1 Towards Inherently Interpretable Machine Learning for Healthcare - A Data-Centric Perspective

As indicated in the introduction, data is of the utmost importance for machine learning. This holds especially true if the aim is to utilize inherently interpretable methods. Therefore, the research I conducted for manuscripts 1 and 2 is focused on creating high-quality data that is specific for the desired prediction task. Furthermore, I explored how high-quality data impacts the possibility to achieve high prediction performances using conceptually simple, i.e., potentially interpretable, models. The prediction tasks investigated in the first two manuscripts are linked by a common issue: missing high-quality training data, i.e., a trustworthy ground truth.

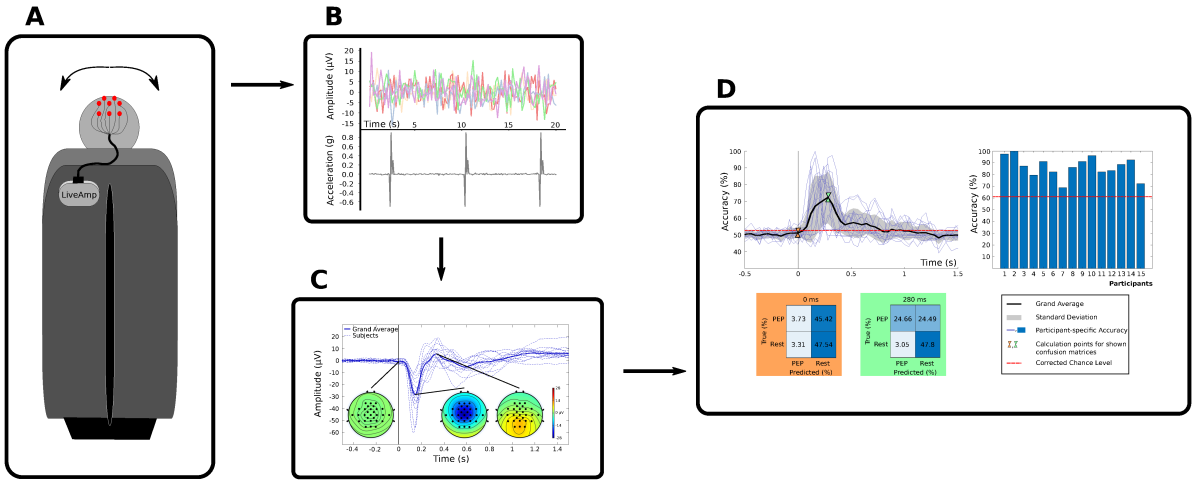

Figure 4.1: Schematic overview of the data collection experiment used to collect high-quality PEP data for classification. **(A)** Participants were seated on a tilting chair with a device that recorded both EEG and acceleration data fixated on the back rest. The chair was tilted by hand. **(B)** The fixation point of the recording device enabled the simultaneous recording of elicited PEPs and movement of the chair. **(C)** Since both data modalities were recorded with the same device, the ensured synchronicity allowed the recovery of well expressed PEPs. **(D)** The data recorded with our experimental setup enabled us to achieve high prediction performance with linear models.

4.1.1 Task-Specific Neural Activity Data (Manuscript 1)

Neural activity recorded with EEG has a notoriously low signal-to-noise ratio [96]. Due to this noisiness, specialized datasets are crucial for the training of models that aim to detect specific activity pattern. In manuscript 1, we tackled the need for such a specialized training set by developing a recording setup that enables us to indicate the occurrence of seated whole-body perturbations of participants without delay. Since the occurrence of perturbation-evoked potentials is highly correlated with the movements of participants that caused these neural patterns, we ensured the same temporal axis for neural and movement data by using the LiveAmp system (BrainProducts, Gilching, Germany) for recording. This device combines an electroencephalogram with an accelerometer. By fixating the LiveAmp onto the tilting chair used for horizontal perturbation, we were able to simultaneously record neural activity and movement of participants (see Figure 4.1 A and B). In contrast to the common use of computational timing flags embedded into frameworks like LabStreamingLayer [97], our setup guarantees a perfect synchronization of perturbation onset and EEG data.

As extensively stated throughout this thesis, interpretable models have to be conceptually simple and, in the ideal case, linear. In order to assess whether or not such a simple model is able to achieve high prediction performance on our PEP dataset, we trained shrinkage linear discriminant analysis (SLDA) classifiers to distinguish between resting EEG and PEPs. SLDA classifiers are often used for the classification of neural activity and are equivalent to Fisher Discriminant Analysis. Since they are based on a linear combination of features using the mean of each class and the common covariance matrix, they are inherently interpretable. We showed

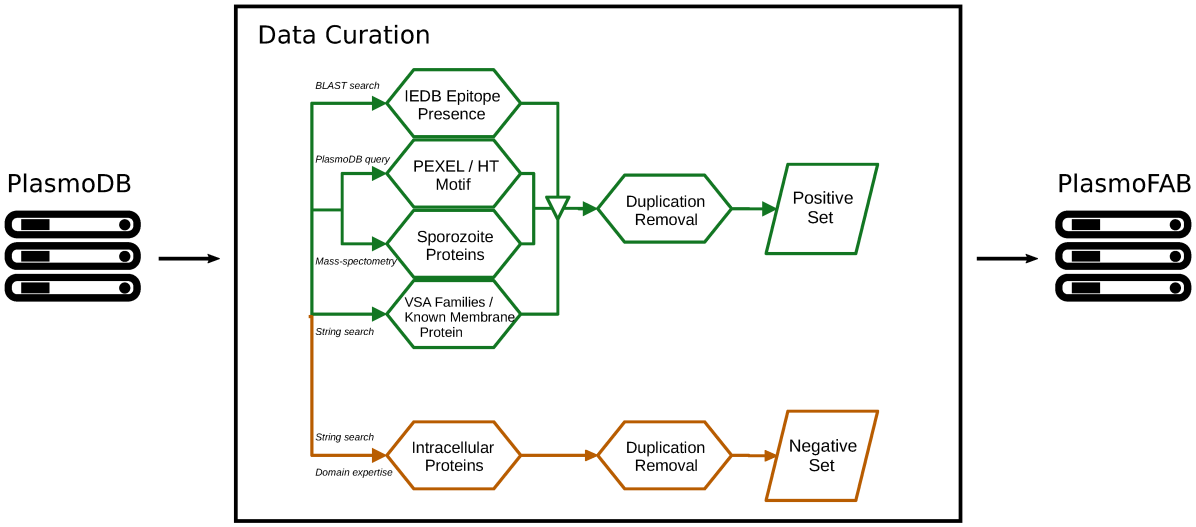

Figure 4.2: Schematic of the data curation process leading to the published benchmark PlasmoFAB. The raw data was extracted from PlasmoDB, a database containing the complete genome of several Plasmodium species. Several pre-processing steps were conducted to create high-quality positive (i.e., protein antigen candidates) and negative (i.e., intracellular proteins) samples. The resulting positive set and negative set were combined into the published benchmark PlasmoFAB.

that even with such a simple predictor, our high-quality data leads to very high prediction performances with an average auROC of 0.93.

We were also interested in different recording scenarios with varying electrode numbers to evaluate the viability of adapting the recording system into real-world applications. The possibility to shrink the number of required electrodes but still achieve a high prediction performance is crucial for integrating such a system into existing hardware. Rehabilitation devices, virtual reality hardware, and other existing solutions are conceptualized to use a minimum of space and the ease of integrating new modules increases with a decreasing module size. We found that even with as little as 5 electrodes, our high-quality data allowed for the training of an SLDA classifier that still reached about 93% accuracy for classifying PEPs. Using a single electrode at channel Cz led to a prediction accuracy of about 87%. These results showed that shrinking the recording module is conceptually feasible.

While the results for the PEP detection task show the potential that specialized data has for enabling machine learning in general and simplifying prediction models in particular, I used the next research project to investigate the potential in terms of a data modality that is more traditionally used in bioinformatics and medical informatics: protein sequences.

4.1.2 Manually Curated Protein Labels (Manuscript 2)

Plasmodium falciparum, the parasite that causes malaria, expresses more than 5300 proteins. Currently, the function of more than half of these proteins is unknown, resulting in the situation

that the majority of *P.falciparum* proteins cannot be included in drug target screenings. Training machine learning models to provide a fast and relatively resource-efficient procedure for identifying protein antigen candidates has the potential of vastly improving the search for an effective drug or vaccine against malaria. The main limitation for employing such ML models is the lack of training data with high-quality labels that indicate whether a sample is a protein antigen candidate. With the second published manuscript, we aimed to close this gap.

The gold standard for *P.falciparum*-specifc proteins to be labeled as potential antibody targets is experimental evidence for the proteins to be visible from the outside of infected host cells. This include transmembrane proteins, surface proteins, membrane-located proteins, and exported proteins. We employed a mixture of established computational algorithms for sequence comparison with literature search, and domain expertise to select a subset of *P.falciparum*-specific protein sequences from the database PlasmoDB. Each protein in the subset fulfilled the condition that it was possible to assign a high-quality (ideally gold standard) label indicating whether it is a protein antigen candidate or not. We published the protein sequences with their corresponding labels as our benchmark PlasmoFAB.

The creation of PlasmoFAB involved a multi-step curation process. First, we identified proteins that contain known epitopes. This term describes the part of an antigen that the host's immune system recognizes. Therefore, the presence of an epitope is a clear identifier for a protein antigen candidate. We used results from the basic local alignment search tool (BLAST, [98]) applied on the Immune Epitope Database* (IEDB) to discover all *P.falciparum*-specific proteins that contain known epitopes. Second, we used PlasmoDB's data fields to identify all proteins that contain a specific sequence motif called Plasmodium exported element (PEXEL) or host targeting (HT). Proteins that contain PEXEL/HT are usually exported into the extracellular space. For the third step, we used experimental evidence published in [99]. In this work, mass-spectrometry was used to identify surface-exposed sporozoite proteins and we included these results in our benchmark. The last step to identify protein antigen candidates involved a combined string and literature search. We performed a string search on PlasmoDB for proteins that belong to specific protein families that fulfill the requirements to be protein antigen candidates. For each hit of our string search, we performed a literature search to verify whether there is experimental evidence. While the previous steps were conducted to identify positive samples for our benchmark, i.e., protein antigen candidates, the next step was conducted to find negative samples, i.e., proteins that do not fulfill the prerequisites to be considered protein antigen candidates. Here we combined a literature search with domain expertise to compile a list of intracellular proteins. Since intracellular proteins cannot leave an infected cell except for specific situations like the burst of an infected cell which only occurs late in the infection cycle, they cannot reliably be targeted by drugs or vaccines. More importantly, these situations usually occur after cell death and, thus, to late for meaningful interventions. Our benchmark was published on Zenodo†.

We used PlasmoFAB to investigate the benefits of specialized ML models over pre-trained prediction services. Prediction services offer pre-trained models that can be easily applied to new datasets. However, they cannot provide any performance guarantees due to issues with

*https://www.iedb.org/

†https://doi.org/10.5281/zenodo.7433086

out-of-distribution generalization. We were able to show that comparatively simple models that are trained on specialized data, i.e., our benchmark PlasmoFAB, vastly outperform the more complex prediction services. Our results clearly show the benefit of creating high-quality data for model training over the use of pre-trained prediction services.

4.1.3 Discussion of Data-Centric Research

With our work published in the article "Perturbation-evoked potentials can be classified from single-trial EEG" we showed that neural activity data specific to the investigated prediction task allows to achieve high performances with linear, thus interpretable, models. Furthermore, specific data enables researchers to investigate different scenarios, e.g., varying recording setups, that are important for a realistic application of these neural activities.

As mentioned in the introduction, the choice of performance metric has substantial implications. The performance of prediction models trained to classify PEPs was evaluated with several different performance measures, including accuracy, true positive/negative rate, and auROC. None of the used performance measures are known to be particularly robust on imbalanced data. However, the use is valid in our case since we carefully balanced the training and validation data used in model performance assessment. The rationale behind the choice of performance measures was that the used values are widely accepted in the neural engineering community and, hence, easily understandable for researchers. Therefore, we decided that the accessibility of our results is more important than a choice of performance metric that enables comparability even if imbalanced data is used in other experiments.

While the experimental setup allowed for the recording of high-quality PEP data, the manual tilting of participants prevented a more detailed analysis of the neural pattern. One of the open questions that we were unable to answer due to the experimental setup (see supplementary material of the original manuscript for more details) is whether the perturbation direction has an influence on the neural pattern. The ability to infer the perturbation direction from the neural activity could be extremely useful in scenarios like aviation and rehabilitation, where this information is needed to compute the system's best possible reaction to a detected perturbation. Substituting the manual tilting for an automatic solution would enable experiments with controlled directional tilting. Another question that we were unable to answer is whether a physical perturbation is needed to elicit a PEP. If only a visual perturbation is needed, this neural activity pattern could be useful for the improvement of virtual reality. A mismatch between visual and vestibular sensory information can lead to dizziness and loss of balance control [87]. If a PEP can be elicited by visual perturbations, detecting this PEP can be used to identify visualizations that are problematic and either replace them with alternatives or try to avoid similar visualizations in the future. These two questions remained open due to the fact that the neural activity data recorded with our experimental procedure was not specific enough for answering them. In other words, the fact that these two questions remained open supports one of the main claims of this thesis: the importance of data's specificity.

With our work published in the article "PlasmoFAB: A Benchmark to Foster Machine Learning for *Plasmodium falciparum* Protein Antigen Candidate Prediction" we show that manually curating labels for protein sequences is key for training machine learning models

that can be used to help advancing healthcare-related research. The most important gain of deploying computational methods for protein prescreening lies in the potential of drastically reducing time- and resource-consuming experimental procedures. Therefore, cost-efficient computational prescreening allows researchers to consider an increased number of proteins for drug and vaccine development.

The main goal of the research that has cumulated into our published benchmark PlasmoFAB was to enable the use of computational methods, in particular machine learning models, for the investigation of proteins that are expressed by the parasite *Plasmodium falciparum*. Although this parasite is one of the most severe threats to human health nowadays, due to the fact that *P.falciparum* causes malaria and the whole genome is extensively sequenced, the function of a majority of *P.falciparum*-specific proteins is still unknown. Furthermore, there were no reliable labels indicating potential protein antigens prior to PlasmoFAB's release. This seriously hindered the deployment of machine learning models to help explore the vast number of proteins with unknown function. We closed this significant gap by publishing PlasmoFAB. However, in the process, we also showed the importance of high-quality data for machine learning in healthcare and, particularly, for interpretable machine learning in healthcare. The performances of pretrained prediction services severely dropped on PlasmoFAB's test data compared to models trained on our specialized dataset. Furthermore, the conceptually simpler interpretable model that we tested (SVM with oligo kernel) was still able to outperform most of the pretrained prediction services.

The focus of the ML models trained in the work described in section 4.1 did not lie on interpretability. Nevertheless, the main result presented in the first two manuscripts is a crucial prerequisite for interpretable machine learning. We showed in manuscript 1 and 2 that a carefully created dataset enables training of prediction models that achieve high performance even if the deployed models were conceptually simple. Although a prediction model's performance is not a sufficient condition to guarantee meaningful interpretations, to achieve a high prediction performance is a necessary condition for the meaningfulness of a calculated interpretation. This becomes obvious when we remember the main purpose of interpretability in machine learning: understanding the cause of a decision. Prediction models with medium to low performance are not able to learn a decision function that captures the pattern or structure within the data that are relevant for the prediction task. If an interpretation for a decision made by such a model is calculated, there is no guarantee that the result of this interpretation contains any information to either advance knowledge about the studied prediction task or help researchers and users validate a decision made by the model. This is true for inherently interpretable and post-hoc interpretation models. Therefore, high prediction performance is a necessary condition for interpretability and the results presented in the first two published manuscripts show that this can be achieved by carefully curating training data as an alternative to deploying overly complex prediction models.

The results presented in the first two published manuscripts demonstrate the importance of high quality data for inherently interpretable machine learning for healthcare. The following section will present the results of the second research direction that I explored towards inherently interpretable machine learning for healthcare.

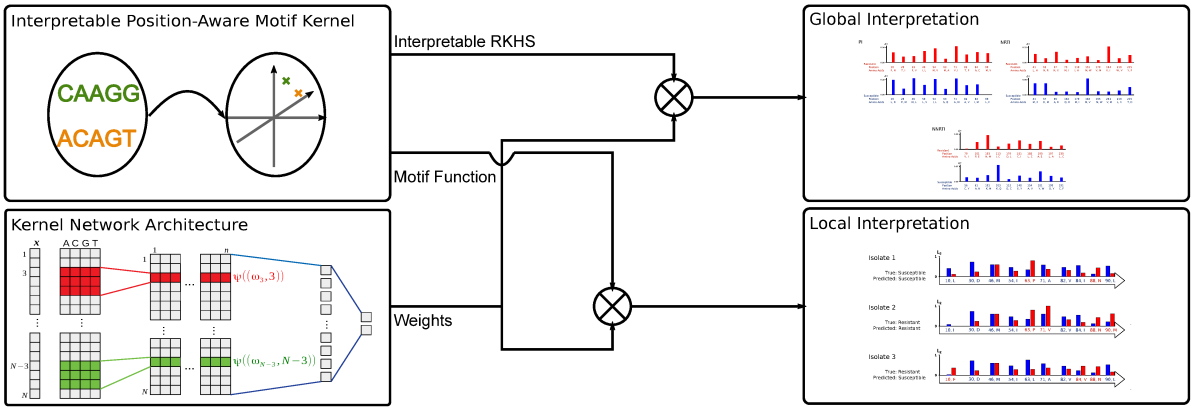

Figure 4.3: We developed a new kernel function called position-aware motif kernel and designed a new neural network architecture that incorporated learning within a subspace of the corresponding RKHS. Our newly developed architecture, called convolutional motif kernel network, results in inherently interpretable prediction models that allows for global and local interpretation without the need for post-hoc techniques.

4.2 Towards Inherently Interpretable Machine Learning for Healthcare - A Model-Centric Perspective

Although data of the highest possible quality is a mandatory prerequisite for interpretable machine learning, the used prediction models also have to be inherently interpretable. Otherwise, an interpretation has to be approximated using post-hoc models. In the following sections, I present my research that was published in the manuscripts 3 and 4. Both manuscripts share a similar research question. How can an inherently interpretable model be developed that achieves state-of-the-art performance on relevant prediction tasks and can be applied on small-, medium-, and large-scale datasets? My research focuses on two of the most commonly used data modalities in bioinformatics and medical informatics: sequence data (manuscript 3) and tabular data (manuscript 4).

4.2.1 Convolutional Motif Kernel Networks (Manuscript 3)

There are two information sources that are important for comparing and distinguishing biological sequences: compositional variability (i.e., the actual sequence of nucleotides or amino acids) and positional variability (i.e., the spatial occurrence of informative substrings). We developed a new kernel function that utilizes both information sources for calculating the kernel value of two biological sequences. This new kernel function, called position-aware motif kernel (PAM), is defined as:

$$K_{\text{PAM}}(x, x') = C \sum_{p=1}^{|x|} \sum_{q=1}^{|x'|} \exp\left(\alpha\left(\omega_p^T \omega_q - k\right) + \frac{\beta}{2\sigma^2}\left(\tilde{p}^T \tilde{q} - 1\right)\right), \tag{4.1}$$

with

$$\tilde{p} = \left(\cos\left(\frac{p}{|x|}\pi\right), \sin\left(\frac{p}{|x|}\pi\right)\right)^T \quad \text{and} \quad \tilde{q} = \left(\cos\left(\frac{q}{|x'|}\pi\right), \sin\left(\frac{q}{|x'|}\pi\right)\right)^T$$

where $|x|$ and $|x'|$ denote the length of the corresponding sequences, ω_p is the motif in sequence x at position p, ω_q is the motif in sequence x' at position q, $\tilde{p}$ denotes the projection of position p onto the upper half of the unit circle, and $\tilde{q}$ denotes the projection of position q onto the upper half of the unit circle. This projection ensures that positions are encoded with unit ℓ_2-norm vectors. This allows to incorporate PAM into a convolutional kernel layer. The parameter α determines the degree of compositional uncertainty, i.e., the influence that mismatching motifs have on the kernel evaluation. The parameter σ determines the degree of positional uncertainty, i.e., the influence that distant motifs have on the kernel evaluation. Finally, the parameter β is used to compensate for the decreased absolute distance that the projection onto the upper half of the unit circle introduces. The constant $C = \sqrt{\frac{\pi^2\sigma^2}{2\alpha\beta}}$ arises from the derivation of the kernel function and the details can be found in the supplement of the original work. We used this new kernel function and a variant of the Nyström method [70, 71, 66] to develop our convolutional motif kernel network (CMKN), an ANN architecture that allows for inherently interpretable learning on biological sequences.

As a first experiment, we created synthetic DNA data with distinct compositional and positional features and investigated whether CMKN models are able to recover the embedded patterns. Our results suggest that CMKN models are able to recover biologically meaningful patterns with high accuracy. Additionally, we used two healthcare-related prediction tasks to evaluate the performance capabilities of our newly developed network architecture: antiretroviral drug resistance prediction of HIV isolates and splice site detection. CMKN models were able to perform similarly to or outperform all state-of-the-art competitors on both prediction tasks. For the antiretroviral drug resistance task, CMKN models successfully learned known drug resistance mutation (DRM) positions from the data. Furthermore, the motifs learned at each position focused mainly on known mutations that are causing drug resistance. On the splice site detection task, CMKN models were able to recover sequence patterns that are associated with real splice sites such as the poly pyrimidine tract before an acceptor site or the AG dimer directly in front of a donor site.

While biological sequences contain valuable information for many healthcare-related prediction tasks, it is well known that an integrated use of several different biological data modalities can provide novel insights that are crucial in investigating diseases and developing novel treatments. Many omics data modalities are stored as tabular data, like gene expression, DNA methylation and others. Therefore, kernel networks have to be utilizable on tabular data to allow the incorporation of more data modalities with valuable information into research projects. The next section describes my research into the development of convolutional kernel networks for tabular omics data.

4.2.2 Convolutional Omics Kernel Networks (Manuscript 4)

As described in the introduction, a kernel function that enables interpretable learning on tabular omics data is the pathway-induced kernel by Manica and colleagues [55]. We developed a convolutional kernel layer that projects input data into a subspace of the kernel's RKHS. Similar to the development of CMKN, a variant of the Nyström method was utilized. However, the expressiveness of PIK (as introduced in section 2.3.1) comes from using several kernel functions in combination, one for each utilized pathway. While previous work used multiple

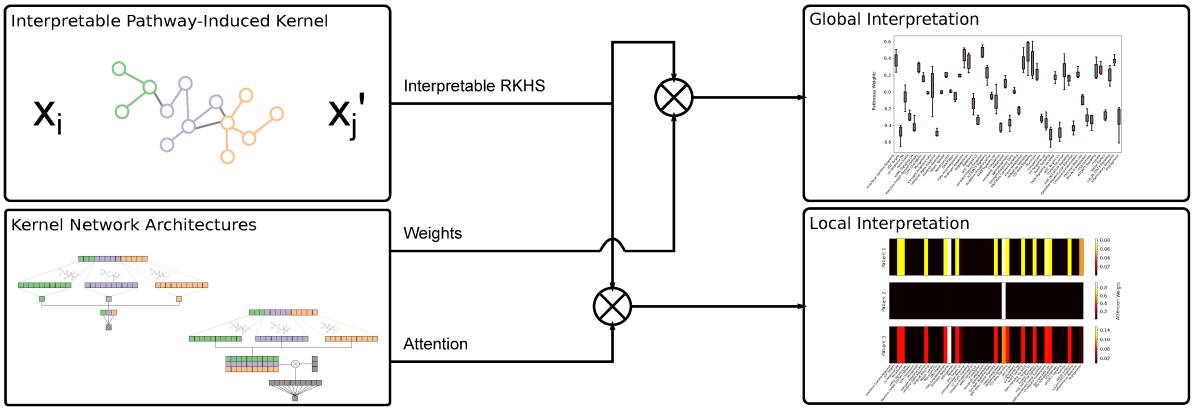

Figure 4.4: We developed a new neural network layer that utilizes the Nyström method to project molecular measurements onto a subspace of the pathway-induced kernel's RKHS. Utilizing this new kernel layer, we showed how different network architectures can be utilized to create inherently interpretable models which either global or local interpretation capabilities.

kernel learning for combining kernel functions, we developed a kernel layer that projects the input into subspaces of several different reproducing kernel Hilbert spaces. Each subspace is representing one specific pathway. A schematic visualization of this layer can be found on the left side of Figure 4.4 or in the corresponding article's Figure IV.1. We call this new network architecture convolutional omics kernel network (COmic).

While the multiple learning approach of previous work allowed for a global interpretation of trained models, our approach enabled us to develop inherently interpretable models with global or local interpretation capabilities. Globally interpretable models are created by combining a max pooling layer and a simple fully-connected layer with linear activation. These two layers combine the projections onto the different RKHS subspaces and calculate a prediction outcome. The weights of the fully-connected layer can be used to provide a global interpretation by indicating the importance each pathway had for the learned prediction task. Locally interpretable models are created by using an attention mechanism. Here, the network computes an attention weight for each subspace projection based on the individual input. These attention weights can be directly used to interpret the influence each pathway has for the prediction made for a specific input, thus they provide a local interpretation. We coined the terms pooling-based COmic for globally interpretable models and attention-based COmic for locally interpretable models.

We tested the prediction performance of COmic models on six different breast cancer cohorts. The task was to classify patients that had a distant metastasis free survival (DMFS) or relapse free survival (RFS) of more than five years. The input was gene expression data. 15 state-of-the-art prediction models that utilize pathway information were used for performance comparison and COmic models were able to either perform similarly to or outperform all competitors in this single-omics prediction task. Furthermore, we were able to show that the global and local interpretation capability of COmic models captured medically relevant patterns within the gene expression data, thus providing evidence for the benefits of our method. We also showed the potential of utilizing COmic models for multi-omics prediction tasks with the METABRIC

cohort. Here, we predicted RFS using gene expression and copy number alteration (CNA) data. Again, COmic models were able to outperform competitors. Additionally, we investigated how well the training of COmic models scales to different dataset sizes and increasing numbers of omics modalities. We were able to show that COmic models can be easily applied to datasets with a vast number of samples and several different omics modalities since the computation time needed for training a Comic model scales linearly in sample count and number of omics modalities.

4.2.3 Discussion of Model-Centric Research

With our work published in the article "Inherently Interpretable Position-Aware Convolutional Motif Kernel Networks for Biological Sequencing Data", we showed that a carefully defined kernel function can be embedded into convolutional kernel networks to create inherently interpretable prediction models for biological sequences. Our method provides end-to-end trainable ANNs that achieve state-of-the-art performance on healthcare-related prediction tasks while enabling researchers and users to compute global and local interpretations without the need for post-hoc models.

In comparison to previously proposed kernel networks, CMKN shows one limitation. The focus of interpretability makes the definition of deeper kernel layers more sophisticated than in previous methods. In other proposed kernel networks, the kernel function used to define a kernel layer can be defined recursively which, in turn, allows for a straightforward creation of deeper kernel layers. However, the resulting RKHS is non-interpretable. We will investigate whether a deeper kernel layer that preserves CMKN's property of being inherently interpretable can be defined in future work. Furthermore, we will investigate how CMKN's architecture can be extended to recover combinations of motifs and positions that are important for the decision. In some prediction tasks based on biological sequences, motifs at specific positions can be non-informative when considered secluded from each other. However, they can be of utmost importance, if considered together. Being able to utilize combinations of motif-position pairs for prediction and interpretation could enhance the performance of CMKN models and their usefulness for knowledge advancement. Another potentially beneficial research direction is the inclusion of phylogenetic information into CMKNs. It is well known that the evolutionary history and relationships among individuals contain valuable information and is reflected within genomic sequences (which is also reflected in amino acid sequences of translated proteins) [100, 101]. Therefore, extending CMKN models to also utilize phylogenetic information for predictions is a promising future research direction.

Another limitation of CMKN comes from the chosen network architecture. Our model is based on a convolutional neural network. While this type of architecture ensures an efficient implementation of the subspace projection, it imposes the strict prerequisite onto input sequences to have the exact same length. Possible solutions to this limitation could be offered by different network architecture, like recurrent networks, and there are published examples on recurrent kernel networks [68]. However, the proposed models did not ensure to be inherently interpretable. Since a vast majority of real-world biological sequence datasets do not fulfill the prerequisite of ensuring that all samples have the same sequence length, a promising future research direction would be to explore the feasibility of inherently interpretable recurrent kernel

networks. If successful, this research would result in an inherently interpretable prediction model that is applicable to a larger number of real-world datasets.

With our work published in the article "COmic: Convolutional Kernel Networks for Interpretable End-to-End Learning on (Multi-)Omics Data", we present an inherently interpretable ANN, trainable using a simple end-to-end learning scheme, that can be applied to tabular omics data. Our model was able to perform similarly to or outperform state-of-the-art competitors on breast cancer survival prediction using different cohorts. Furthermore, we showed that our network architecture allows to create global and local interpretations without the need for post-hoc models.

The main obstacle to applying COmic models to new datasets is the creation of appropriate Laplacian matrices. Currently, these Laplacians have to be created manually. This provides the benefit that users can tailor these matrices to incorporate specific prior knowledge that they decide to be optimal for the prediction task at hand. However, manually creating Laplacians can be time-consuming and requires some knowledge in graph theory. This could prevent researchers in the biological or medical field to utilize COmic. A possible solution to this problem is learning the Laplacian matrices directly from the input data. Several publications show that inferring Laplacians from graph signal data is possible [102, 103]. Recently, it was also shown that specialized neural networks can learn Laplacians from graph signal data using gradient-based end-to-end learning [104]. However, the learning of graphs is computationally highly expensive and only feasible for small graphs. Realistic biological graphs usually consist of several hundred nodes and further research is required to investigate whether the previously described methods can be scaled to larger graphs. Incorporating Laplacian inference into COmic models could allow users to only specify a subset of the input features for which a Laplacian is then directly learned. There are two main advantages in extending the COmic functionality in this way. First, users' workload would be significantly reduced since predefined Laplacians are not required. Second, users can use this functionality to investigate different feature combinations, e.g., different sets of genes in gene expression data, in order to find new graphs or pathways that provide additional or new knowledge about relevant endpoints, like diseases.

The proposed methods, CMKN and COmic, result in inherently interpretable models, as shown with the results published in the corresponding manuscripts. In their seminal review about interpretability and explainability in AI, Arrieta and colleagues introduced three levels of transparency that an inherently interpretable model can fulfill [74]. From highest to lowest, these levels of transparency are *simulatability*, *decomposability*, and *algorithmic transparency*. Both of the proposed models fulfill the requirements to be at least in the second highest level of transparency, *decomposability*, if the interpretation is aimed at an audience of domain experts. The projection onto an intelligible RKHS together with the use of strictly linear layers enables domain experts to understand and explain all parts of the resulting model. Furthermore, it provides them with the ability to understand the behavior of a CMKN/COmic model. However, under certain circumstances, CMKN/COmic models can be categorized within the highest level of transparency, *simulatability*. For a model to fall within this category, a human has to be able to simulate it or strictly think about it [74]. For CMKN models, this property is fulfilled if the number of anchor points is kept low and the length of biological sequences is manageable. For COmic models, this property is fulfilled if the number of anchor points and

the number of pathways is kept low.

The results published in manuscript 3 and 4 indicate that a correctly chosen kernel function creates inherently interpretable kernel networks with state-of-the-art performance capabilities. However, kernel networks offer additional benefits. On the one hand, they allow to scale kernel learning to large-scale datasets. We showed with COmic that inherently interpretable kernel networks can be utilized on datasets with hundreds of thousands of samples with relative ease. Consequently, our proposed models mitigate the scalability issue of traditional kernel methods and provide a feasible solution to apply kernel learning within the big data regime. As previously mentioned, the size of biological and medical datasets are expected to vastly increase in the near future which renders scalable models an important necessity. On the other hand, kernel networks offer the possibility to make certain kernels computationally feasible. This point refers mainly to our introduced position-aware motif kernel (PAM). While there is a similarity between PAM and the oligo kernel (introduced by Meinicke and colleagues [53]), the latter had to find a trade-off between computational feasibility and expressiveness due to the fact that the oligo kernel was meant to be used with traditional kernel methods like SVMs. Therefore, the oligo motif was limited to discrete k-mers which ensured an efficient computation of needed Gram matrices. For PAM, there is no efficient way to implement the computation of the Gram matrix rather than a greedy approach due to the use of motifs. This leads to serious drops in computational performance with increasing sample numbers, motif sizes, and sequence lengths. However, kernel networks do not compute Gram matrices and can push the expressiveness of kernels by mostly ignoring the trade-off between computational feasibility and expressiveness.

While biological sequences and tabular omics data are the main data modalities used in personalized medicine, there are additional modalities used in healthcare that are not covered by the presented prediction models. Three of the more prominent data modalities in healthcare are written doctoral nodes, a non-standardized textural data modality used by healthcare practitioners to transfer diagnosis and treatment information about patients to other professionals, time-series data containing information about changes in the health status of patients and treatment responses, and medical imagery data produced by procedures like computed tomography (CT) scans or magnetic resonance imagery (MRI). Inherently interpretable models for these data modalities need to be developed to enable a full integration of inherently interpretable machine learning into healthcare. Therefore, future research should investigate the possibility to develop novel models that allow for inherently interpretable learning on data modalities that are not yet covered by the presented methods but provide important information for healthcare.

4.3 Integrated Discussion

With this thesis, I present the research I conducted during my doctoral studies. The common theme of my research was to explore the applicability of inherently interpretable machine learning in healthcare. Decision-making in healthcare and, in particular, in precision medicine depends on the utilization of high-dimensional and multi-modal data generated by individual patients. Machine learning has a huge potential for these two research fields due to the ability to detect patterns within high-dimensional data and correlate these patterns with

specified outcomes. However, there is a significant limitation to fully utilizing machine learning for knowledge advancement in the mentioned fields. ML experts can build highly complex models to optimize predictions on high-dimensional data but usually lack the expertise to fully conceptualize medical research. On the other hand, healthcare experts, like medical doctors, offer the needed domain expertise but lack the training to conceptualize and, hence, understand predictions made by ML models. I argue that inherently interpretable models can act as a bridge to bring together ML experts and healthcare experts and have the potential to positively impact the interdisciplinary research that is needed in healthcare and precision medicine. I approached the viability of inherently interpretable machine learning for healthcare from two different research directions.

The first part of my research explored the impact of high-quality data on utilizing prediction models for healthcare-related tasks and the possibility for achieving high performances with conceptually simple models, an important prerequisite for the viable application of inherently interpretable models. One argument that is often presented against inherently interpretable models is that they lack the expressiveness of black-box models due to their innately simpler concept. However, my results show that the argumentation against conceptually simpler models can be countered by putting more effort into increased data quality. If the used data is highly complex and lacks specificity for the investigated prediction task, increasing a model's complexity can positively impact the achieved prediction performance. Since the higher complexity of data usually results in more complex patterns that are correlated with the targeted end point, the higher expressiveness of conceptually complex black-box models can lead to an improvement of learning these complex patterns. However, on the task of classifying neural activity associated with perturbations, we showed that carefully ensuring a high specificity of the training data allows for substantial prediction performances even with simple linear models. Additionally, we showed that an interpretable model trained on specialized protein data outperforms complex pre-trained black-boxes. For the sake of completeness, I have to point out that the best performance on the *Plasmodium falciparum*-specific protein antigen classification task was achieved by combining a black box feature encoder, namely the language model ProtT5, with an inherently interpretable prediction model. Nevertheless, the overall results of my data-centric research towards inherently interpretable machine learning in healthcare show the importance of putting a focus on data quality for the goal of making inherently interpretable models viable in healthcare. It is worth noting that the importance of data quality is widely recognized in biological and medical science and, based on the results of my work, I strongly support current movements in the machine learning research community that the importance of data quality should be one of the future focuses of research in machine learning.

The efforts in my data-centric research were focused on the *specificity* aspect of the datasets, using the definition of this aspect that was provided in the introduction. Since the manual curation and validation of data points and labels requires a lot of labor, my colleagues and I were able to create small- to medium-scale datasets with the resources available to us. However, that limited the possibility to investigate the impact of the second aspect of data that I defined in the introduction: *completeness*. Since data is only an abstraction of reality, sufficient samples are needed to ensure that a research subject can be thoroughly investigated. This holds especially true in healthcare, where there are numerous ways to stratify patients including ethnicity, genetic variations, environmental exposures, age, societal differences, and many

more. The described aspect of completeness was not part of my research since creating a large number of high-quality data is impossible for a single researcher or research group. This requires large community efforts and there are already such efforts ongoing like the international cancer genome consortium (ICGC, https://dcc.icgc.org/). However, I am confident that such high-quality and large datasets will become more widely available in the future. To fully benefit from these community efforts, inherently interpretable models need to be scalable to large-scale datasets. This motivated the second part of the research I conducted towards the overarching goal of my thesis: model-centric research.

The model-centric part of my research focused on investigating the possibility to make kernel networks inherently interpretable. In my opinion, artificial neural networks are among the most interesting types of prediction models currently used in machine learning. What motivates me to make this statement is the fact that the underlying idea of ANNs is incredible simple, propagating activation through a weighted network and adjusting the weights based on a computed loss, yet they can become arbitrarily expressive and, in theory, can learn any function. Due to their expressiveness being based on the used architecture, ANNs are a highly flexible type of prediction model that can be easily adapted and optimized for different data modalities and prediction tasks. Furthermore, due to the possibility of massively parallelizing the training procedure, ANNs can be applied to any dataset from small-scale data with a few hundred samples to large-scale data with hundreds of thousands of samples. However, the expressiveness of neural networks comes with the cost of becoming highly complex deep learners that are virtually impossible to understand for humans, i.e., black boxes. The advances in combining ANNs with kernel methods in recent years [65, 66, 105] showed that it might be possible to overcome the black-box nature of ANNs without sacrificing their expressiveness, although previous work did not try to create inherently interpretable prediction models. With the results achieved by the two kernel networks proposed in my research, CMKN and COmic, we have shown that neural networks can have high prediction performances while being inherently interpretable. That opens the door to applying inherently interpretable machine learning models to large-scale biological and medical datasets.

However, before I start discussing the benefits that inherently interpretable models can bring to the research in biology, medicine, and healthcare, I would like to discuss a few benefits that inherently interpretable models can offer to machine learning projects in general. As mentioned in the introduction, the common practice in ML is to use the same performance metrics during the training of prediction models and for evaluating the training success, i.e., applying trained models to test splits or new data. One might argue that iterative, gradient-based training procedures use a loss function instead of a performance metric during the training, but model selection, an important part of the training process, is usually based on performance metrics not the loss function. The resulting situation is that the measure of success becomes a target during training. And, as stated by Goodhart's law, a targeted measure ceases to be a good measure. I would like to argue that inherently interpretable models can offer a solution to this issue. Being able to directly access the decision surface and, thus, interpret the reason behind a prediction made by a model allows researchers and users to use prior knowledge to evaluate the validity of a model's decision making process. For example, if a model is trained to make a prediction based on gene expression data and the best performing model only uses genes that are known to be unrelated to the predicted endpoint, the validity of the model has to be questioned even though it achieved the best performance. One possible

explanation for this situation could be that the model is relying on spurious correlations. An issue that would be impossible to detect with looking at the performance measure alone. This example is just a simplified thought experiment but it shows the potentially positive impact that inherently interpretable models can have. However, like all expertise-driven systems, this type of evaluation is far more labor-intensive than simply comparing performance measures. A significant amount of domain expertise is needed to evaluate the decision process of an inherently-interpretable model and experts have to judge whether the use of features that are unknown to be related to the endpoint points towards the exploitation of spurious correlations or points towards gaps in the currently available knowledge about the investigated subject.

Another problem that can be mitigated by inherently interpretable models is resource consumption in terms of needed computation time for model training. Nowadays, it is well known that deep learning models require significant amounts of energy for training. Using GPUs for training a single instance of a base version of BERT, an NLP model that has to be considered small in modern standards, produces approximately 652 kg CO_2eq of emissions [106]. And usually deployed training procedures train hundreds or even thousands of model instances. As exhaustively discussed during this thesis, one prerequisite for making models inherently interpretable is decreasing the complexity of models. This decreased complexity is accompanied by a decrease of model size, thus, the resources needed to train inherently interpretable models are less than for bigger, more complex models. Another, less obvious, aspect for reduced resource consumption during training of inherently interpretable models is hyperparameter optimization. Usually, these hyperparameters are optimized by training models with different hyperparameter combinations, e.g., using a manual, grid, or random search [107, 108, 109]. These methods require to retrain models for each tested combination, which immensely increases the resource consumption of the training procedure. However, as shown in manuscripts 3 and 4, the hyperparameters of my developed inherently interpretable models have meaning within the data domain and, thus, can be selected based on prior knowledge. This possibility drastically reduces the number of hyperparameter combinations that have to be tested. We show in manuscript 3 and 4 that selecting hyperparameters based on domain expertise actually leads to prediction models that outperform competitors. This demonstrates that inherently interpretable models can drastically reduce resource consumption by providing a resource-efficient (basically free) method of hyperparameter optimization. I view this second benefit as the more crucial one. The current state of our world with the ongoing climate catastrophe demands fundamental changes about the way we live, how our economy works, and how we consume [110, 111, 112]. Reducing resource consumption is not an option, it is a necessity if we want to soften the impact of the climate catastrophe and retain the earth as a livable environment. While there is an increasing amount of efforts and events dedicated to this purpose, e.g., the ICT4S conference series[‡], reducing resource consumption should become a major consideration in the mainstream ML research and inherently interpretable models have the potential to help with that.

Apart from the general benefits for machine learning, inherently interpretable models have the potential to significantly impact the medical field. Given the fact that biological data consists of several different modalities and most medical questions can only be answered by combining information from more than one data modality, prediction models can help refining

[‡]https://conf.researchr.org/series/ict4s

the stratification process of patients, a procedure that is often referred to as personalized medicine. However, black-box models hide their decision process behind highly complex mathematical projections that deny domain experts any sense of agency when working with them. Either they trust the model's prediction or they have to revert to established diagnosis approaches for patient stratification to regain a sense of agency. This creates situations of unequal power, where healthcare practitioners are denied the authority in their own field of expertise. Additionally, it leads to the disregard of knowledge that was acquired over many years and sometimes even decades. While this transfer of power in everyday medical practice is concerning, the potential negative impact of unequal power on interdisciplinary research projects should also not be underestimated. In healthcare, the detailed mechanism of action for many diseases and, especially, treatments is not completely understood yet. For example, there exists a group of genes that are known to be involved in the processing of pharmaceuticals. These ADME (**a**bsorption, **d**istribution, **m**etabolism, and **e**xcretion) genes play a vital role in many treatments like chemotherapy but which genes belong to the ADME group and the precise effect of these genes on different treatments remain open research questions [113, 114, 115]. Black boxes could hinder knowledge advancement in these cases. As an example, utilizing black boxes to predict the optimal dose of a pharmaceutical for the chemotherapy of a patient might reduce the negative impacts of this treatment option and simultaneously increase the reduction of cancer cells, but it obscures the decision boundary used to solve this problem from domain experts. Maybe the black box's decision boundary could point towards mechanisms of action that are not yet known. This information could help domain experts to improve existing treatments or develop new treatments with the potential to mitigate negative impacts and improving the recovery prognosis for patients. One might argue that this can be solved by using post-hoc interpretability methods, an approach that is usually called XAI (explainable artificial intelligence). However, there are severe shortcomings to the application of XAI methods. These models can only approximate the real decision surface of the used prediction model [85]. Furthermore, they are optimized to yield an explanation that "makes sense" in the eye of the person who is training the explanation model [85, 116]. Since these are usually machine learning experts with limited domain expertise, there is a real danger that explanations that could point towards new mechanisms of action are discarded in favor of explanations that fit more with the limited understanding of the ML expert. Inherently interpretable models do not suffer from these limitations. They provide direct access to their decision boundary. Furthermore, the interpretation of a decision is provided within the data domain on which the model is trained as shown by the results presented in the two manuscripts described in the model-centric part of my research. Domain experts can use their own knowledge to directly evaluate the decision made by an inherently interpretable model and investigate whether the correlations used by the prediction model point towards unknown mechanisms of action that could advance the current knowledge about treatments and diseases. Therefore, inherently interpretable models have the potential to positively impact interdisciplinary research in general and medical research in particular.

I used this integrated discussion to share my educated opinion about the benefits that inherently interpretable prediction models offer to machine learning research, especially in interdisciplinary projects. This leaves me with the opportunity to sketch out my vision for future research conducted on the topic of this thesis, which is inherently interpretable machine learning for healthcare. Unfortunately, making prediction models inherently interpretable remains a

niche research direction within the wider ML community. However, I do hope that increasing efforts will be invested into further researching the potential of inherently interpretable models due to the fact that the awareness towards potentially severe flaws of solely relying on black-box models gets more and more traction among ML researchers and political stakeholders [83, 117, 118, 119, 120]. The results achieved by the inherently interpretable kernel networks presented throughout this thesis indicate that these models can achieve high scalability and expressiveness. Research questions that remain open include investigating the possibility for deep kernel networks that remain inherently interpretable or the exploration of new kernel functions that are computationally infeasible for traditional kernel methods but can be used with kernel networks and incorporate currently unused domain expertise. I do believe that we have only scratched the surface of the real potential of inherently interpretable models and I am excited to see what the future has in store for this research direction. However, there is one important issue that cannot be solved by traditional machine learning, not even with inherently interpretable models, and I call this issue *abstraction of causality*. What I mean by this term is the fact that scientists are usually not only interested in solving a prediction problem but want to explore the causal structures of the world that allow a certain prediction to be solvable. Interpreting the cause of a decision made by a prediction model does not reveal information about causality in the data, i.e., in the world, but only causality within the model. In other words, the result is an abstraction of the world's causality by the model's causality. There is a scientific discipline[§] that tries to infer causal structures within the world, called causal inference [121, 122, 123]. In recent years, there are novel developments to incorporate causal inference into the machine learning framework [124, 125, 126]. I see the possibility to solve the *abstraction of causality* issue with machine learning models that are able to truly infer causal structures within the data. However, this research direction is only at the beginning and currently faces several challenges that need to be overcome before causal models can be generally used in application scenarios. One of these challenges is that causal inference and, thus, causal machine learning is based on graphical models. Currently, the utilized graphs tend to be on the smaller size for real-world applications in healthcare [127] and a general issue arises from the need to validate that derived graphs are able to capture the causal structure in the data. There is no universally agreed upon solution for such a validation procedure which often results in controversial discussions whether a published causal model is valid. However, I do believe that these issues will be resolved in the future given the previous mentioned fact that research about combining causal inference and machine learning is still at the very beginning and gained increased attention from researchers in recent years. Once causal machine learning becomes more widely accepted, I see immense potential for causal models to be applied to prediction problems in healthcare and advancing the knowledge about diseases and treatments.

[§]With the term "scientific discipline" I refer to work that is conducted with regard to the scientific method.

5 Conclusion

Nowadays, medical routines in diagnosis and treatment often rely on an overly simplified stratification of patients. This is necessary due to the fact that humans are notoriously bad in conceptualizing high-dimensional spaces. Therefore, doctors would be overwhelmed if presented with the raw biological data that each individual patient produces and resort to using easy-to-access and easy-to-understand data like age, weight, sex, body temperature, electrocardiograms, or described symptoms when diagnosing and treating patients. Furthermore, the advancement of knowledge about diseases and potential cures (drugs or vaccines) depends on a complete understanding of the biological status of patients. In my opinion, machine learning has the potential to provide substantial benefits to medical research and healthcare due to the ability to find patterns in high-dimensional spaces and relate these patterns to outcome variables.

However, ML experts often lack the knowledge to fully conceptualize the complexity of medicine and healthcare. I argue that significant advancement can only be achieved through an interdisciplinary effort. Inherently interpretable machine learning can help to realize such an expert-in-the-loop scenario with a balanced power dynamic between the involved disciplines. My research shows that inherently interpretable models can achieve state-of-the-art prediction performance on healthcare-related tasks and provide biologically meaningful interpretations. Furthermore, inherently interpretable models allow for an interpretation of decisions within the data domain, thereby allowing domain experts to use them for evaluating predictions and advance knowledge, and provide direct access to their decision surface. This possibility to directly access the learned decision surface mitigates the issue of ambiguous interpretations due to hyperparameter choices in post-hoc models. Therefore, the use of inherently interpretable models offers the additional benefit of a possible solution for decreasing the chance of discarding interpretations that could lead to novel insights.

So: Why are you using models that you do not understand to investigate something that you do not understand?

6 Appendix

Publications contributing to this doctoral thesis, as listed in Chapter 1, are included in the following appendix. The corresponding citations can be found in Chapter 1. All publications are printed as published except some minor template changes.

License Information

I Perturbation-Evoked Potentials can be classified from single-trial EEG

Jonas C. Ditz Andreas Schwarz Gernot R. Müller-Putz

Abstract

Objective: Loss of balance control can have serious consequences on interaction between humans and machines as well as the general well-being of humans. Perceived balance perturbations are always accompanied by a specific cortical activation, the so-called perturbation-evoked potential (PEP). In this study, we investigate the possibility to classify PEPs from ongoing EEG.
Approach: 15 healthy subjects were exposed to seated whole-body perturbations. Each participant performed 120 trials; they were rapidly tilted to the right and left, 60 times respectively.
Main Results: We achieved classification accuracies of more than 85% between PEPs and rest EEG using a window-based classification approach. Different window lengths and electrode layouts were compared. We were able to achieve excellent classification performance (87.6 ± 8.0% accuracy) by using a short window length of 200 ms and a minimal electrode layout consisting of only the Cz electrode. The peak classification accuracy coincides in time with the strongest component of PEPs, called N1.
Significance: We showed that PEPs can be discriminated against ongoing EEG with high accuracy. These findings can contribute to the development of a system that can detect balance perturbations online.

I.1 Introduction

The sense of balance is crucial for humans in their everyday routine. Standing and walking are not possible without it. Every human being learns balance control during early childhood and the loss of balance control always leads to uncomfortable, often potentially dangerous situations. The possibility to compensate for loss of balance control can alleviate harmful consequences and, therefore, vastly improve human experience. In gait rehabilitation, exoskeletons are used to support patients during rehabilitation sessions [86]. However, with the ability to compensate for loss of balance control, the support provided by an exoskeleton can be limited to an on demand state. Thereby, the independence of patients can be increased for better rehabilitation. The system takes control from the patient if it is needed to prevent falling.

In virtual reality (VR), the conflict between sensory and vestibular information can lead to different physiological effects, inter alia, postural instability [87]. If the system is able to detect postural instability, i.e. perceived balance perturbation, alternative visualization protocols as well as emergency shut downs of the visual environment can be put into action as soon as they are needed. Nevertheless, a reliable and potent detection method for the loss of balance has to be found.

Electroencephalography (EEG) studies found a specific activity pattern that was elicit as a response to a balance perturbation [128, 129, 130, 131, 132]. This cortical activity, called perturbation-evoked potential (PEP), consists of four distinguishable parts. After an initial small positive wave (P1), a large negative deflection (N1) follows. The third and fourth part

is again a positive wave (P2) followed by a negative wave (N2). The last two parts are often collectively referred to as late perturbation-evoked response (PER). Timings of the different parts usually reported by researchers are 30 - 90 ms after perturbation onset for P1, 80 - 160 ms after perturbation onset for N1, and 200 - 400 ms after perturbation onset for late PERs [133].

PEPs have been thoroughly investigated on the neurophysiological level, linking them to cortico-cortical transfer processes [134], error-potentials (mismatch between actual and expected position) [132], and compensatory motor planning processes [135]. Regardless of the underlying neural processes, loss of balance control is always accompanied by a PEP (especially the N1 component is always reproducible) independent of the mode of perturbation [133]. Therefore, a system that can reliably detect changes in state of mind (i.e. occurrences of specific neural activation patterns) can be used for the detection of PEPs.

Control over a computer or machine by solely using one's mind is the main goal of research in the field of Brain-computer interfaces (BCIs) [136, 137, 138]. While there has been a strong focus on using BCIs for controlling assistive devices [139, 140], BCIs can improve the interaction between humans and machines in the context of human-machine interaction (HMI). These so-called passive BCIs (pBCI) do not provide active control to users; moreover, they monitor their state of mind and detect changes in the state of mind of users [141, 142]. Studies have shown that implicit information about the state of mind of a user can be found in distinct brain patterns. Working memory load can be detected by monitoring oscillatory power in theta band over frontal-midline electrodes [143]. Error-related potentials are specific activity patterns that are elicited when a user makes or perceives an error and are used to compensate for erroneous interactions [144, 145, 146]. Another potential that is used in pBCIs is the Bereitschaftspotential (BP), which precedes spontaneous movements [147, 148]. Due to their ability to detect changes in user's mental state, pBCIs are a promising tool for detecting perceived balance perturbation.

In this study, we investigate whether PEPs can be autonomously discriminated from ongoing EEG. Fifteen healthy participants were exposed to seated whole-body perturbations. Each participant performed 120 trials; they were 60 times rapidly tilted to the right and 60 times rapidly tilted to the left. We developed a method that can decode PEPs from ongoing EEG recordings. Additionally, we evaluated parameters imperative for boosting PEP classification for existing pBCI systems such as different window lengths and smaller electrode layouts. Finally, we constructed an offline scenario to test our PEP detection method.

I.2 Materials and Methods

I.2.1 Participants

Fifteen healthy participants (6 female, 9 male) took part in the study. Participants were between 19 and 57 years old with an average age of 26.7 $\pm$ 9.4 years. All participants had normal or corrected-to-normal vision and were without any known medical condition. The study was approved by the ethics committee of the Medical University of Graz. All participants gave written informed consent and received monetary compensation for their efforts.

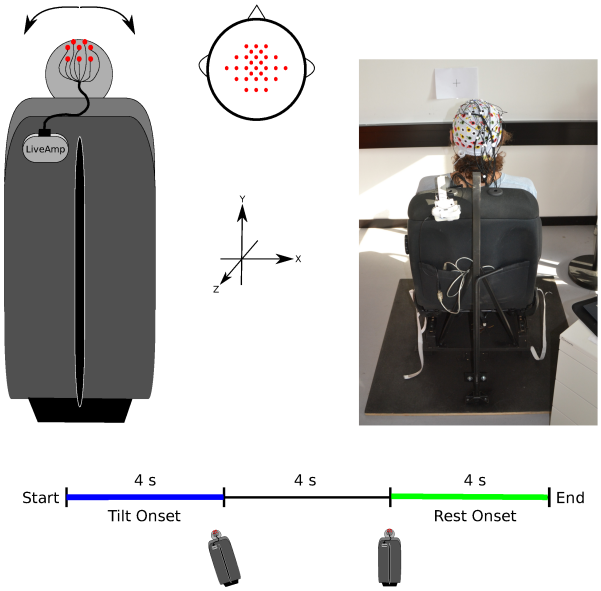

Figure I.1: A schematic of the experimental setup is shown on the left side. The axes shown left from the schematic indicate the alignment of the accelerometer. In the center, the electrode layout used for recording is displayed. A picture of the actual chair used for perturbation with a participant sitting in the chair can be seen on the right side. The structure of trials is shown on the bottom. The perturbation onset was randomly set within the first four seconds of each trial (marked in blue). The chair stayed in the tilted position until the start of the break eight seconds after the start of the trial. The break (marked in green) had a duration of four seconds. Within this period, the artificial rest onset is located.

I.2.2 Hardware and data acquisition

We recorded EEG from 29 active electrodes using a LiveAmp system (BrainProducts, Gilching, Germany). Electrodes were distributed over frontal, central, and parietal areas at positions F1, Fz, F2, FFC1h, FFC2h, FC1, FC3, FCz, FC2, FC4, FCC1h, FCC2h, C5, C3, C2, Cz, C2, C4, C6, CCP1h, CCP2h, CP3, CP1, CPz, CP2, CP4, P1, Pz, and P2. Additionally, we used three electrodes to record electro-ocular activity (EOG) and placed them above the nasion and below the outer canthi of the eyes. The ground electrode was placed at position AFz and all electrodes were referenced to the left mastoid. EEG was sampled at 500 Hz. All signals were pre-filtered using a 3rd order sinc low-pass at 131Hz (cutoff). We also recorded the participant-specific electrode positions using an ultrasound-based ELPOS system manufactured by Zebris (Zebris Medical GmbH, Isny im Allgäu, Germany).

I.2.3 Experimental task

All measurements took place in the BCI-Lab of the Institute of Neural Engineering at Graz University of Technology. Participants were equipped with an electrode cap and seated in a custom-built tilting chair. Using a mechanical tilting system, we were able to tilt the chair 5° to the left or to the right. The amplifier was fixed to the back of the chair and a fixation cross

was put on the wall in front of the chair (see Figure I.1). Participants were asked to take a comfortable position in the chair and rest their arms on their legs in order to reduce muscle tension in the arms and shoulders. We further instructed them to stay relaxed and fixate the cross in front of them during the whole experiment. In the beginning of the experiment, two perturbations (one perturbation to the right, one to the left) were performed to familiarize them with the task at hand. Subsequently, each trial had the following structure: Within the first four seconds of each trial at a randomly chosen time point, participants were rapidly tilted either to the left or the right (based on a random generator). The chair was tilted manually using a mechanical lever. The chair stayed tilted until 8 seconds after trial start. Thereafter, the chair was put back into the neutral position for an inter-trial interval of 4 seconds. In this way, each participant experienced 60 perturbations to the left and 60 to the right in random order (120 perturbations in total). Each event during the experiment was indicated by a marker. The synchronization of marker data and amplifier was realized using LabStreamingLayer (LSL) [97].

I.2.4 Perturbation onset detection

We recorded the perturbation onset for each trial using the intrinsic accelerometer (3 axes) of the LiveAmp, which was fixed on the backside (upper left corner) of the tilting chair (see Figure I.1). We applied thresholding on the first derivative of the abscissa (x axis in Figure I.1) to acquire trial based perturbation onsets.

I.2.5 Data preprocessing

In order to detect artefact-contaminated trials and exclude them from further analysis, we performed statistical tests. For this purpose, data was band-pass filtered between 0.3 and 35 Hz (zero-phase Butterworth filter, 4th order) and three different statistical parameters were considered for the rejection of artifact-tainted trials. First, we performed an amplitude threshold rejection removing all trials with an amplitude that exceeded ± 125 μV. Afterwards, we tested trials for an abnormal joint probability and an abnormal kurtosis. The rejection threshold was four times the standard deviation (STD) for both tests. 13.8% of the trials were rejected on average. The approach used for outlier rejection does not need additionally recorded channels and is well tested in different BCI scenarios [149, 150]. After artifact rejection, we band-pass filtered the raw EEG between 0.3 and 10 Hz using a causal 4th order Butterworth filter.

I.2.6 Perturbation-evoked potential

We epoched the EEG from -0.5s to 1.5s with respect to the perturbation onset acquired from the accelerometer. Additionally, we acquired rest trials with a length of 2s 4.5s prior to the perturbation onset from -4.5s to -2.5s. Therefore, our time region of interest (tROI) had a duration of 2s. For each participant, we calculated the average over all trials for each condition (perturbation, rest) as well as the 95% confidence interval using nonparametric t-percentile bootstrap statistics ($\alpha = 0.05$) [151]. Additionally, to account for statistically significant

Table I.1: Different window lengths used for binary classification of PEPs with the corresponding number of features. For each window the subject-specific offset with the highest discriminability between rest trials and PEP trials was deduced in a calibration phase.

Window size	**Number of features per channel**	**Number of features per trail**	**Trail-to-feature ratio**
1 sample	1	29	8.28
200 ms	6	174	1.38
400 ms	11	319	0.75
600 ms	16	464	0.52

differences between conditions, we performed the nonparametric Wilcoxon Rank Sum test ($\alpha = 0.01$) on each time point. We corrected for multiple comparisons (with n = 1250 comparisons) using the Bonferroni correction*.

I.2.7 Binary single-trial classification

We used a two-phase window-based approach comparable to the method used in [152, 153] for the classification of PEPs. A schematic of the classification pipeline can be found in the supplementary material (Figure S1). In the beginning, we resampled our data to 25 Hz in order to save computational effort and split our data into two sets: calibration set (containing two thirds of the data) and test set (containing the remaining third of the data). In order to simulate online behavior, we took the trials that were recorded first for the calibration set and the trials that were recorded last for the test set. In the first phase, called the calibration phase, we trained shrinkage linear discriminant analysis (sLDA) classifiers [154, 155] using only trials in our calibration set. In order to train the classifiers, we performed a 10 times 5-fold cross-validation. In each fold, we moved a window over our tROI in steps of 40 ms, i.e. we trained a classifier every 40 ms (2000ms / 40ms + 1 = 51 classifiers for the whole tROI). For each time point, features were extracted by taking amplitude values of each electrode in steps of 40 ms, i.e. the number of features varies depending on the used window length (see Table I.1 for an overview of tested window sizes). The number of features of, for example, a 200 ms window is 200ms / 40ms + 1 = 6, where we add one since the first sample in the window is included as a feature. We separated the calibration set into training and validation set in each cross-validation fold and trained the classifiers using only trials from the training set. Afterwards, we evaluated each classifier using the validation set. After cross-validation, we calculated the mean validation accuracy for each time point and used the best performing classifier of the time point with the highest mean validation accuracy for the second phase.

In the second phase, called the test phase, we extracted features from the unseen test set similar as in the calibration phase. We applied the trained classification model to the features of the test set. Accuracies stated in this work are always test accuracies, i.e. the classification

*A bonferroni correction for 1250 comparisons using a significance level of $\alpha = 0.01$ leads to significant p values at $p < 0.000008$.

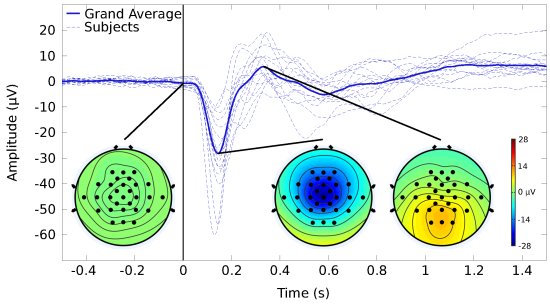

Figure I.2: Grand average of perturbation trials at channel Cz. Dashed lines show the average for each of the participants. The vertical line at time point 0 ms indicates the perturbation onset. Topographical plots show the scalp distribution of the grand average at t = 0 ms, t = 146 ms, and t = 320 ms. The N1 component (topographical plot at t = 146 ms) is distributed over frontal, central, and parietal areas with the center at Cz. The P2 component (topographical plot at t = 320 ms) is distributed over parietal areas with the center at Pz.

performance of our classifier on the test data.

I.2.8 Classification parameter optimization

We tested four different window lengths. The parameter for each of the used window lengths can be found in table 1. Additionally, we tested four different electrode layouts:

- minimal electrode layout (1 channel): Cz
- small electrode layout (5 channels): FCz, C1, Cz, C2, and CPz
- medium electrode layout (15 channels): FC1, FCz, FC2, FCCh1, FCCh2, C3, C1, Cz, C2, C4, CCPh1, CCPh2, CP1, CPz, and CP2
- full electrode layout (29 channels)

For each setup (combination of window length and electrode layout), the classification accuracy was calculated as described above. A two-way anova was conducted to test for a significant effect of the two independent variables (window length, electrode layout) on the classification performance. We tested the effect on peak accuracy as well as peak latency. Each independent variable had four levels (1 sample, 200 ms, 400 ms, and 600 ms for window length; minimal, small, medium, and full for electrode layout). Afterwards, we compared the classification performance of different window lengths as well as the classification performance of different electrode layouts.

I.3 Results

I.3.1 Perturbation-evoked potential

Figure I.2 shows the grand average PEP for channel Cz (solid blue line) as well as individual PEPs of each participant (dashed blue lines). In perturbation trials, the grand average shows a strong negative shift (N1 component of PEPs) starting shortly after perturbation onset and

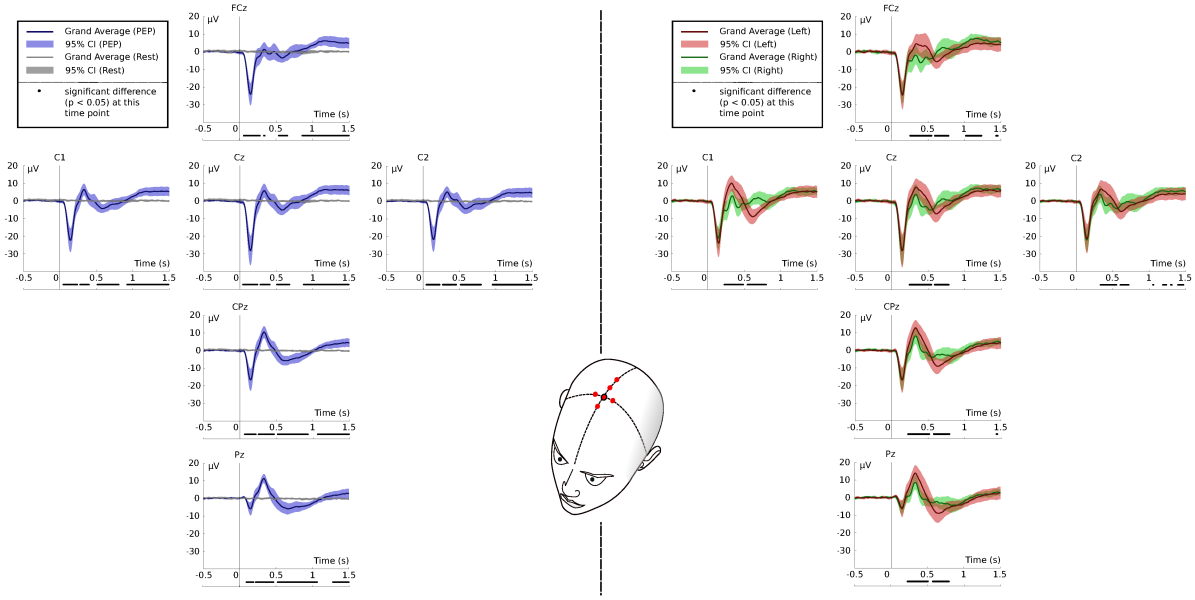

Figure I.3: Grand average of perturbation (blue) and rest (grey) trials. Furthermore, the grand average of "left perturbation" trials is shown in red and the grand average of "right perturbation" trials is shown in green. Shaded areas show the 95% confidence interval (CI) over all subjects. Significantly different time points (Wilcoxon test, $\alpha < 0.01$, with Bonferroni correction for n = 1250 comparisons) between perturbation and rest are marked beneath the plot. Similarly, significantly different time points between left and right trials are also marked beneath the plot.

peaking 144 ± 9 ms after perturbation onset with an average amplitude of -28.3 ± 14.5 μV. The negative peak is followed by a strong positive rebound, the P2 component of PEPs, that peaked on average 330 ± 30 ms after perturbation onset with an amplitude of 12.2 ± 4.1 μV. Neither the P1 nor the N2 component were clearly detectable in the grand average. We also show the grand average on a topographical level for timepoints t = 0 ms (perturbation onset), t = 144 ms (N1 peak) and t = 330 ms (P2 peak): At perturbation onset (0 ms) on average, no visible derivation from the baseline is found. While the peak of the N1 component was centered around Cz and distributed over frontal, central, and parietal areas, the P2 peak was mainly distributed over parietal areas and centered around Pz.

Figure I.3 shows the participant based grand average for channels positioned at FCz, Cz, CPz and Pz as well as positions C1 and C2. The left side compares the perturbation condition versus the rest condition. The black bars beneath each plot indicates significantly different time intervals between rest condition and perturbation condition (Wilcoxon test, $\alpha < 0.01$, with Bonferroni correction for n = 1250 comparisons). The right side of Figure I.3 compares trials where the participant was tilted to the left with trials where the participant was tilted to the right. Similar to the left side, significantly different time intervals between both conditions are indicated beneath each plot.

Grand average of both perturbation and rest conditions show significant differences in morphology starting around 40 ms after perturbation onset. At C1, Cz, C2, CPz and Pz,

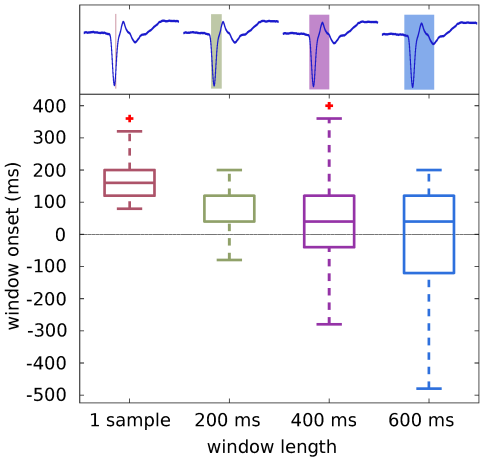

Figure I.4: Onsets of training windows for all window sizes and each subject. Onsets are shown in milliseconds relative to perturbation onset. The perturbation onset (t = 0 ms) is indicated by a dashed line. Position of the average training window for each of the used window sizes is displayed above the box plot.

both N1 and P2 components are significantly different to the rest condition. At FCz, only the N1 component is significantly different compared with the rest condition. Between 500 and 2000 ms after perturbation onset, a small negative peak followed by a positive rebound can be observed for all four channels. The difference of this behavior compared to the rest condition is significant for all channels displayed on the left side of Figure I.3. The N1 component showed no significant difference between right and left perturbations, as can be seen in Figure I.3. The P2 component had a slightly higher amplitude for left tilting trials (16.7 ± 5.3 μV at Pz) compared to trials where the subject was tilted to the right (10.3 ± 5.6 μV at Pz). To test the effect of tilting direction on the amplitude of the P2 component, a one-way between subjects ANOVA was conducted using the peak amplitude and latency at electrode Pz. A significant effect of tilting direction on P2 amplitude at the $p < 0.05$ level was found for the two conditions (left vs. right) [$F(1, 28) = 11.83$, $p = 0.002$]. Additionally, the effect of tilting direction on P2 latency was tested using a one-way between subjects ANOVA. In this case, no significant effect at the $p < 0.05$ level was found for the two conditions [$F(1, 28) = 0.25$, $p = 0.62$].

I.3.2 Binary single-trial classification

We performed binary classification of perturbation against rest trials. The participant-specific chance level for binary classification as well as the chance level for the grand average were calculated using an adjusted Wald interval ($\alpha = 0.05$) [156, 157]. The result was corrected for multiple comparisons (with n = 76 comparisons) using a Bonferroni correction. The participant-specific chance level is 61.05% and the grand average chance level is 52.84%. All classification accuracies reported in this section were achieved on test data.

The calibration phase was used to investigate the time window of maximal discriminability between rest and perturbation conditions. Figure I.4 shows the participant-specific onsets of training windows relative to perturbation onset in the box plot. The average position of the

Table I.2: Results of binary classification for different window sizes. Peak accuracy and latency with respect to perturbation onset are shown as mean and standard deviation of the individual participant accuracies and for the grand average

Window size	**Peak accuracy (%, participant-specific)**	**Peak accuracy (%, grand average)**	**Latency (ms, relative to perturbation onset, participant-specific)**	**Latency (ms, relative to perturbation onset, grand average)**
1 sample	93.4 ± 5.4	81.6	221.3 ± 108.9	120
200 ms	95.1 ± 5.8	83.5	277.3 ± 76.3	280
400 ms	95.3 ± 5.5	78.4	434.7 ± 157.8	440
600 ms	95.7 ± 4.4	76.2	536.0 ± 165.5	640

different sized training windows is shown above the box plot. For a training window consisting of a single sample, the window onset was 184.7 ± 80.4 ms (mean ± STD) after perturbation onset. The average window onset for a training window with a length of 200 ms was 82 ± 74.2 ms (mean ± STD) after perturbation onset. When a window length of 400 ms was chosen for training, the window onset was on average 58.7 ± 74.2 ms (mean ± STD) after perturbation onset. An average window onset of 17.3 ± 152.2 ms (mean ± STD) before perturbation onset was calculated for a training window with a size of 600 ms.

Both independent variables (window length and electrode layout) had a statistically significant effect on the peak accuracy at the 0.05 level. The result of the two-way ANOVA for the window length was [$F(3, 224) = 1.96$, $p = 0.121$]. For the electrode layout the main effect yielded [$F(3, 224) = 15.04$, $p < 0.001$]. The interaction of both independent variables did not have a statistically significant effect on peak accuracy: [$F(9, 224) = 0.06$, $p = 0.999$]. We performed Tukey's honest significant difference criterion to compensate for multiple comparisons (Tukey, 1949). The test showed that the smallest (1 sample) and largest (600 ms) window lengths were significantly different at the 0.05 level. For the second independent variable, the test showed a significant difference at the 0.05 level between the minimal layout and all other layouts. We then took a more detailed look at the classification differences between the tested window lengths. The full electrode layout (29 channels) was used for the comparison. In Figure I.5(a) the classification accuracy of all four window lengths for each time point of the tROI are plotted on the left side while the right side shows boxplots of the participant-specific peak accuracies for each condition. Above chance level classification was possible for all four window lengths. However, peak accuracy shifted away from perturbation onset with increasing size of the window. Table I.2 summarizes the results for each window. The high inter-participant variability of peak latency led to the difference between grand average peak accuracy and the mean of participant-specific peak accuracy.

Furthermore, we investigated how a reduction of EEG channels impacts classification performance. We tested four different layouts: the full layout using all 29 recorded channels, a medium sized layout with 15 channels, a small layout with 5 channels, and a minimal layout with one channel. Based on our previous results, we used a window length of 200 ms for the comparison. Figure I.5 (b) shows the grand average classification performance for each time point of tROI

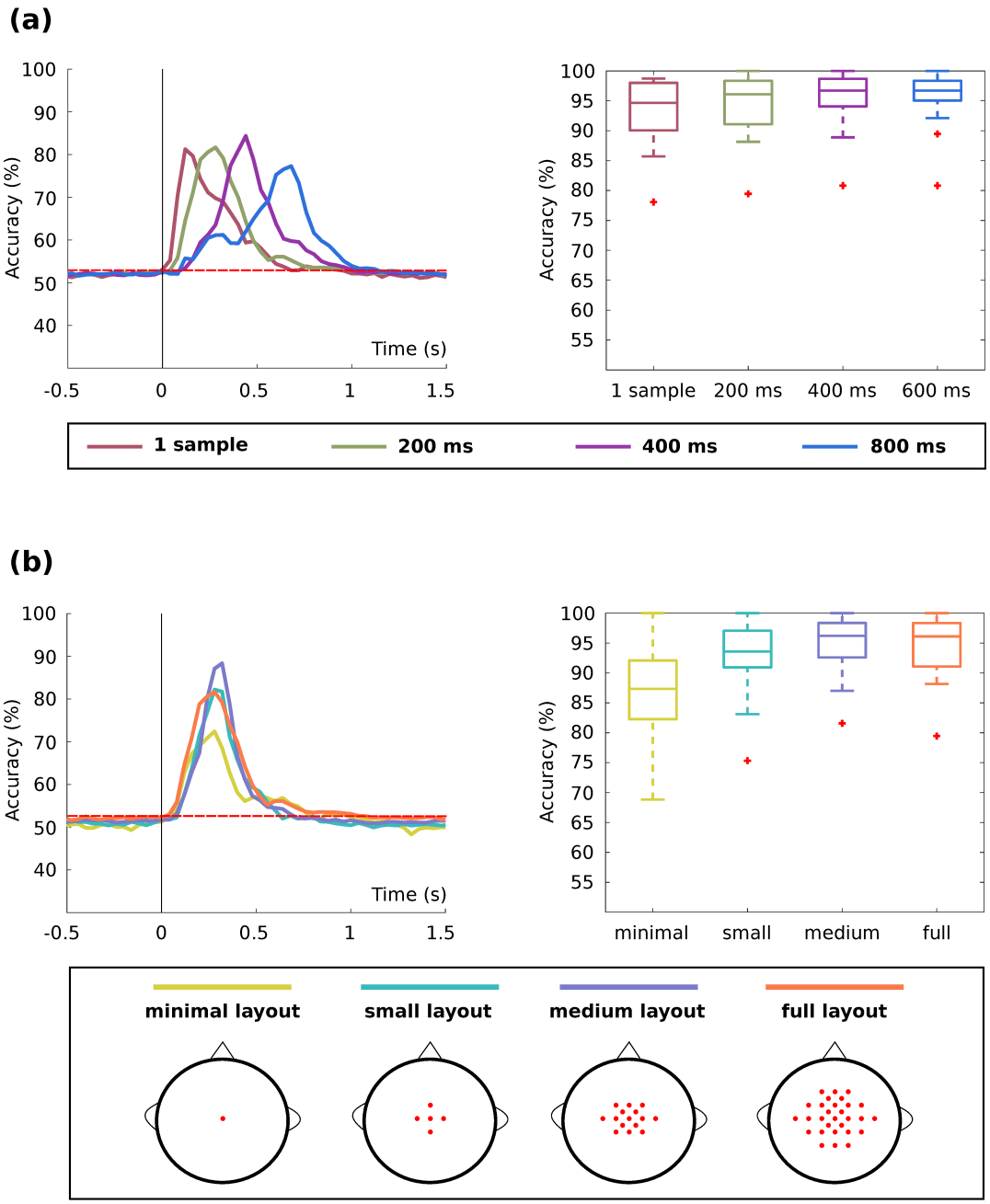

Figure I.5: **(a)** Classification results for different window sizes using the full layout. The left side shows grand average classification accuracies over the course of the tROI. The box plot on the right side shows subject-specific peak accuracies for each of the tested window lengths. Outliers are marked with red crosses. **(b)** Classification results for different layouts using a 200 ms window. The left side shows grand average classification accuracies over the course of the tROI. The box plot on the right side shows participant-specific peak accuracies for each of the tested electrode layouts. Outliers are marked with red crosses.

Table I.3: Results of binary classification for different layouts using a window size of 200 ms. Peak accuracy and latency with respect to perturbation onset are shown as mean and standard deviation of the individual participant accuracies and for the grand average.

Layout	**Peak accuracy (%, participant-specific)**	**Peak accuracy (%, grand average)**	**Latency (ms, relative to perturbation onset, participant-specific)**	**Latency (ms, relative to perturbation onset, grand average)**
minimal (1 channel)	87.6 ± 8.0	71.5	240.0 ± 78.6	280
small (5 channels)	93.1 ± 5.9	81.2	282.7 ± 76.3	320
medium (15 channels)	94.4 ± 5.6	82.7	282.7 ± 73.2	280
full (29 channels)	95.1 ± 5.8	83.5	277.3 ± 76.3	280

on the left side. The right side displays a box plot of peak accuracies of all subjects for each layout. All electrode layouts achieved above chance level performance. The results for each electrode layout are summarized in Table I.3. Differences between participant-specific and grand average peak accuracy occurred due to variations of peak latency between participants.

Based on our previous results, we decided to use the minimal layout with a window length of 200 ms for further analysis. Figure I.6 shows classification results for each of the subjects and the grand average on the left side. Shaded areas indicate the standard deviation (STD) of the grand average. The dashed red line marks the grand average chance level. The confusion matrices shown on the bottom of the left side were calculated for the perturbation onset at 0 ms and for the time point of the maximal accuracy peak at 280 ms. The peak accuracy for each of the participants is shown on the right side of Figure I.6. Participant-specific chance level is indicated by a dashed red line. The grand average classification accuracy exceeded chance level performance between 40 ms and 780 ms after perturbation onset with a peak accuracy of 71.5% correctly classified trials. Since we evaluated the performance of a classifier every 40 ms, we were able to calculate confusion matrices for different time points of the perturbation. In Figure I.6, we show the confusion matrix calculated at perturbation onset (0 ms) and at the maximal accuracy peak (280 ms). At perturbation onset, the classifier achieved a true positive rate (TPR) of 10.8% and a true negative rate (TNR) of 93.3%. The Cohen-α coefficient for this time point was 0.364. At the accuracy peak 280 ms after perturbation onset, the classifier achieved a TPR of 49.7% and a TNR of 92.5%. Here, the Cohen-α coefficient was 0.593. TPR, TNR, and Cohen-α coefficient were calculated using the confusion matrices of the grand average classification result.

Each participant exceeded participant-specific chance level classification performance. The maximum accuracy was achieved 240.0 ± 78.6 ms (mean ± STD) after perturbation onset. On average, the participant-specific classification accuracy peaked at 87.6 ± 8.0% (mean ± STD). The difference between grand average peak accuracy and the participant-specific results occurred due to inter-participant variability of peak time which can be seen on the left side of Figure I.6. For this reason, we calculated TPR, TNR and Cohen-α coefficient for each participant at the participant-specific peak times. Participants achieved a TPR of 81.5 ± 12.2% (mean ± STD) and a TNR of 93.5 ± 6.4% (mean ± STD) on average. The average

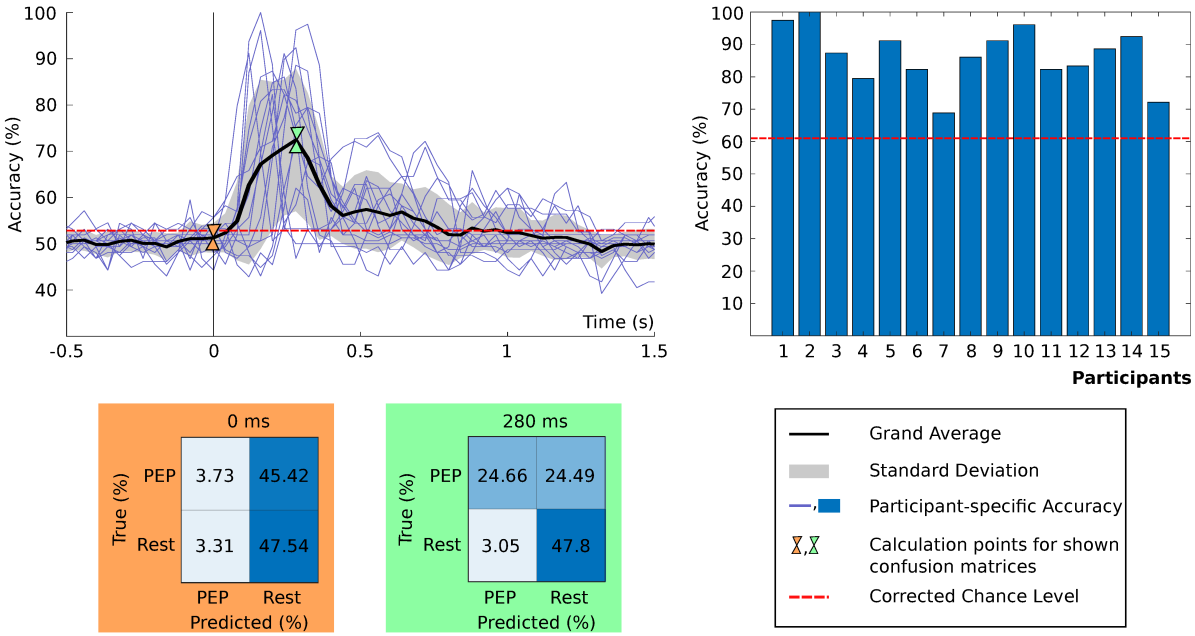

Figure I.6: Classification performance of binary classification using a window length of 200 ms and the minimal electrode layout. The upper left plot shows the classification performance of individual subjects over the course of the tROI in blue. The black curve displays the grand average result with the standard deviation depicted by grey areas. The dashed red line indicates the grand average chance level. The calculation time points for the subjacent confusion matrices are indicated by colored marker. The peak accuracy achieved by each of the participants is shown in the upper right plot. Participant-specific chance level is indicated by a dashed red line.

Cohen-α coefficient was 0.804 ± 0.112 (mean $\pm$ STD).

In order to get a better understanding of our classifier's behaviour, we calculated receiver operating characteristic (ROC) curves for three different time points in our tROI: At perturbation onset (0 ms), at the time point of grand average peak accuracy (280 ms), and by combining time points of participant-specific peak accuracy. The curves are shown in Figure I.7. Furthermore, we calculated the area under the ROC curve (AUROC) for all three curves. AUROC was 0.54 at perturbation onset while it reached 0.73 at grand average peak accuracy. When combining the best classification performances of all participants, AUROC reached 0.93.

I.4 Discussion

In this study, we successfully decoded perturbation evoked potentials on a single trial basis in a controlled laboratory environment. We further identified tuning parameters such as the length of the feature window and the number of EEG channels used to provide configurations for out-of-the-lab use. Post hoc offline analysis showed that even when using only one EEG channel and a feature window of 200ms, participant-specific performances consistently exceeded 70% accuracy, on average peaking at more than 85%. Underlying EEG correlates show significant

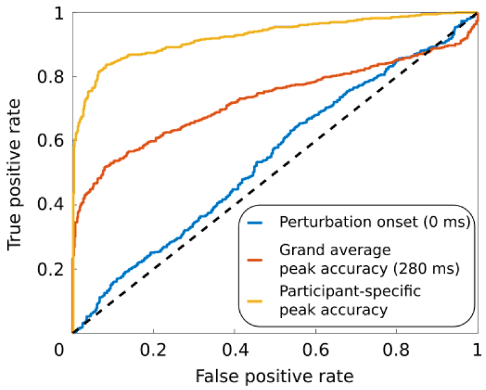

Figure I.7: ROC curves of the binary classifier at different time points. Curves were calculated at perturbation onset (0 ms), the grand average peak accuracy (280 ms), and the combined time points of participant-specific peak accuracy.

differences between the perturbation condition and the rest condition which correspond to the time window used for feature extraction of the best performing classifiers.

I.4.1 Perturbation-evoked potential

We were able to elicit PEPs using the paradigm and experimental setup described in the method section of this work. Our findings were in agreement with typical morphology of PEPs reported in [128, 134, 130, 131, 158].

We observed an average amplitude of -24.8 μV for the N1 component of PEPs elicit during our experiment. On average, the N1 component peaked with a latency of 148.5 ms relative to perturbation onset. The N1 was spread around frontal, fronto-central, and central areas with a maximal peak at Cz. These findings are in accordance with results published in previous studies [159, 160, 158, 161].

The late components of PEPs, called P2 and N2, can be mainly found between 200 and 400 ms after perturbation onset [133]. We were able to identify one of those components (P2) while the other component (N2) was not identifiable. On average, the P2 peaked 328 ms after perturbation onset with an amplitude of 12.1 μV. This finding agrees with the results of previously conducted studies involving seated perturbations. Mochizuki and colleagues reported similar latencies for identified P2 components in participants that were exposed to postural perturbations while seated [162]. However, they found higher amplitudes (23.66 $\pm$ 6.21 μV). This difference could be explained by the difference in task design between our study and Mochizuki et al.: While participants in their experiment had a vertical pole that they were holding during perturbations which induced a compensatory arm reaction, no movement reaction was induced in our experiment. A decreasing P2 amplitude due to lack of balance reaction would however disagree with findings by Quant and colleagues. They reported that the P2 amplitude is higher in passive trials, i.e. trials that do not evoke a compensatory balance reaction [163]. This disagreement between their and our results could occur due to different types of perturbation. In our experiment, participants were exposed to whole-body

perturbations while the PEPs in Quant's experiment were elicited using dislocation of the feet.

The P1 component of PEPs was not distinguishable from background EEG for all of the recorded participants. This finding is consistent with P1 responses observed in previously conducted studies involving seated perturbation tasks [131, 163, 162]. The P1 has a small amplitude which can result in difficulties separating this early PEP component from background EEG. Since the perturbation was induced by hand, there were small mechanical differences in the perturbation movement between the trials. Together with the small amplitude, these variations could have led to the absence of a P1 response in PEPs of single participants due to the averaging of trials (Figure I.2).

The comparison between PEPs elicited during right perturbation trials and left perturbation trials showed that the P2 component had a higher amplitude during left trials. Evaluating the accelerometer data showed a difference in the perturbation speed/acceleration between left and right trials (see Figure S3 in the supplementary material). Furthermore, we surmise that the amplitude difference is partly caused due to our referencing with only one electrode at the left mastoid. To confirm this theory, we performed a subsequent reanalysis of the comparison between left and right trials and applied a common average reference (CAR) [164, 165]. The difference in P2 amplitude at Pz between the two executed perturbation directions was not detectable after re-referencing our recorded data using this spatial filtering approach (P2 amplitude at Pz left trials: 11.3 ± 3.7 μV; P2 amplitude at Pz right trials: 11.1 ± 3.2 μV; see Figure S2 in the supplementary material). The result of our reanalysis together with the accelerometer data supports the assumption of an artificial difference between left and right trials due to several confounders. This agrees with findings published in literature, where no effect of motor or sensory information [166] or psychological factors [167] on late PEP components was found. Since there is a high probability that the difference between PEPs elicited during left trials and PEPs elicited during right trials is not founded in neural activity, we did not attempt to train a classifier for the discrimination of left and right trials.

I.4.2 Binary single-trial classification

The result of binary classification shows that a PEP can be discriminated from ongoing resting EEG with an accuracy of above 80%. To the best of our knowledge, there are no other studies with the goal of classifying PEPs. We used a calibration approach to detect the time interval in which the discriminability of perturbation and rest trials is maximal. Our results show that this time interval is congruent with the occurrence of the N1 component around 140 ms after the perturbation onset. All of the four tested window lengths are centered around this EEG correlate. Since the PEP has a participant-specific latency, this centering around the EEG correlate explains the window onset variability shown in Figure I.5. This finding supports our hypothesis that a PEP can be decoded by the electrical properties of this activity pattern.

The reason why we chose the window sizes used in this work is that 1 sample serves as a minimal example while 200 ms is enough to fully envelope the N1 component. 400 ms and 600 ms are used to analyze how classification performance is affected by longer time windows. Our investigation of different window sizes for the classification showed that there are no significant differences between the peak classification performances of the four tested window lengths.

This is not surprising for the three longer window sizes (200 ms, 400 ms, and 600 ms) since the classifier mainly uses information provided by the N1 component for the detection of a PEP and all three window lengths are big enough to fully contain this component. It is interesting that a single sample is already enough to achieve a classification performance similar to those achieved with longer windows. This can be explained by the high amplitude of the N1 component, which is big enough to encode sufficient information in a single sample. However, the usage of only a single sample can prove to be problematic since such a restriction in terms of time makes the classifier vulnerable to artifacts. On the other hand, an increase in window size means that the static delay introduced by the window-based classification approach will also increase. The 200-ms-window achieved peak accuracies around 100 ms after the amplitude peak of the N1 component in contrast to a delay of around 300 ms for the 400-ms-window and a delay of around 400 ms for the 600-ms-window. Our results suggest that a window length of 200 ms is a good trade off that allows for a classification that is stable and robust against artifacts and introduces a reasonable delay. We used the full layout to compare different window lengths in order to utilize all information that was available to us when assessing the difference in classification performance of different windows.

Since our previous results show that there is no significant effect of window length on classification peak accuracy, we decided to use the 200 ms window for the comparison of different layouts because this window length achieves a good tradeoff between static delay and robustness. Four layouts were considered: the full layout of 29 electrodes, a medium layout with 15 electrodes, a small layout consisting of 5 electrodes, and a minimal layout with only 1 electrode. While the full, medium, and small layout performed without significant difference reaching peak classification accuracies around 94% on average, the minimal layout reached slightly lower accuracy peaks with a maximal classification performance of 87.6% on average around 240 ms after the perturbation onset. This finding agrees with the localization of the PEP N1. Since the component is distributed over frontal, central, and parietal areas with the peak amplitude located at Cz, the removal of channels that are far away from Cz does not change the classification accuracy and suggests that no additional discriminating information can be found in these channels. This hypothesis is supported by the high accuracies of well above 80% (peak) achieved with the minimal layout that uses only the Cz electrode.

To analyze our classifier in more detail, we looked at the classification results using a 200 ms window length and the minimal electrode layout. We chose these parameters due to the findings discussed above. We found that all participants were able to reach a peak classification accuracy above 70%. We calculated confusion matrices at different timepoints to judge the behaviour of our classifier. The confusion matrix calculated at perturbation onset (0 ms) indicated that almost all trials were classified as rest trials (see Figure I.6). This behavior is expected since all trials look like rest trials at perturbation onset (flat EEG). Therefore, classifying all trials as rest trials is the correct behavior at perturbation onset and our classifier showed this behavior. The confusion matrix showed that only half of the PEP trials were correctly classified at grand average peak accuracy. However, there was a high variety of classification peak latency, i.e. for many participants we did not take the best performing time point when calculating the confusion matrix. For this reason, we calculated confusion matrices for each participant at the specific peak time. These matrices showed much better results (TPR: 81.5 $\pm$ 12.2%; TNR: 93.5 $\pm$ 6.4%; Cohen-α: 0.804 $\pm$ 0.112). Furthermore, we calculated AUROC values at different time points to get a better performance measurement of our classifier at different points in our tROI.

At perturbation onset, we calculated an AUROC value of 0.54. Since every trial looked like a rest trial at perturbation onset, as previously mentioned, we expected the classifier to fail at this time point and the AUROC value supports our expectation. An AUROC value of 0.73 indicates that our classifier only reached fair performance at grand average peak accuracy. Again, this behavior is not unexpected since there was a high inter-participant variability of peak accuracy. When only the best performances are considered, by combining participant-specific time points of peak accuracy, we found that our classifier reached an classification performance with an AUROC value of 0.93.

I.4.3 Limitations and future work

We prepared resting trials by using a virtual onset between perturbations. Participants were instructed to stay as relaxed as possible and did not perform a mental or physical task during these periods. In a real-world scenario, however, PEPs have to be discriminated from ongoing EEG during mentally or physically demanding tasks like flying an airplane or doing physical therapy for the purpose of rehabilitation.

Another limitation of selecting resting trials in this way is that subjects are constantly prepared for being tilted. Although we randomized perturbation onsets to some degree, this anticipation of the perturbation cannot be completely prevented. Due to the manual tilting during trials, the acceleration showed a systematic divergence between the two tilting conditions. This is unfortunately a confounder, preventing further investigations between left and right perturbations. However, the ability to distinguish tilts to the left from tilts to the right is of great importance for several fields such as application of exoskeleton. Therefore, future experimental setups should incorporate controlled tilting procedures to validly investigate different tilt directions.

The use of a window-based classification approach will introduce a static delay in an online scenario. Although other online BCI systems do not suffer severely from that delay [168, 169], a BCI that is supposed to detect loss of balance control works under strong time restrictions. One example would be a patient in rehabilitation therapy to restore lost walking functionality. The patient is walking with minimal support but is attached to a system that prevents falling if a loss of balance control is detected. Such a system has to react in an instant to prevent the patient from falling and hurting themselves. Future research should address the delay introduced due to the window-based classification approach and investigate whether the duration of that delay could be problematic in real-world scenarios.

The classifier used for discriminating PEP and rest condition, namely sLDA, has one important restriction: the performance depends on the estimated covariance matrix. However, the stability of this estimation decreases with an increase of the size of the feature space (curse of dimensionality; see [155]). The usage of a shrinkage algorithm already addresses this problem but the use of further dimension reduction techniques could improve the classification in terms of performance and stability. Successfully implemented methods are e.g. principal component analysis (PCA) [170], sequential forward selection (SFS) [171], or smoothing with a moving average filter [172].

We have shown that classification can be performed with high accuracy using only a small number of channels (93.1% with five channels; 87.6% with one channel). This result is crucial for real-world scenarios since size is an important factor in integrating new parts into an existing system. Furthermore, companies that are specialized in EEG hardware are investigating into shrinking the size of their hardware during the last years to improve mobility. Although the developments in the hardware section are promising for PEP-detection systems, a thoroughly online study has to be conducted to really determine the viability of such a system. In such an online study, one has to investigate not only the reliability of PEP detection but also the false positive rate (TPR). The TPR should be well below one per minute since compensating loss of balance control usually restricts the user and such restrictions should only be put in place if they are necessary. Finally, tests have to be performed with actual use cases of a system that compensates loss of balance, e.g. with patients during gait rehabilitation, while using a VR headset, or during actual flights.

I.5 Conclusion

In this study we showed that perturbation-evoked potentials can be robustly discriminated from ongoing EEG using a linear classification approach. Furthermore, we show that this discrimination can be achieved using only a few electrodes for recording. Our findings are a first step to make information about the user's balance control state accessible for computers and machines. This would enable a machine to react to perceived balance perturbations which could improve interactions between humans and machines. An improvement of this interaction is important to enhance rehabilitation medicine or VR experience.

Acknowledgement

The authors would like to thank Dietmar Josef Schäfauer for his help in planning and constructing the tilting chair used to perturb participants. Supported by TU Graz Open Access Publishing Fund.

II PlasmoFAB: A Benchmark to Foster Machine Learning for *Plasmodium falciparum* Protein Antigen Candidate Prediction

Jonas C. Ditz* **Jacqueline Wistuba-Hamprecht*** **Timo Maier** **Rolf Fendel** **Nico Pfeifer** **Bernhard Reuter**

Abstract

Motivation: Machine learning methods can be used to support scientific discovery in healthcare-related research fields. However, these methods can only be reliably used if they can be trained on high-quality and curated datasets. Currently, no such dataset for the exploration of *Plasmodium falciparum* protein antigen candidates exists. The parasite *Plasmodium falciparum* causes the infectious disease malaria. Thus, identifying potential antigens is of utmost importance for the development of antimalarial drugs and vaccines. Since exploring antigen candidates experimentally is an expensive and time-consuming process, applying machine learning methods to support this process has the potential to accelerate the development of drugs and vaccines, which are needed for fighting and controlling malaria.
Results: We developed *PlasmoFAB*, a curated benchmark that can be used to train machine learning methods for the exploration of *Plasmodium falciparum* protein antigen candidates. We combined an extensive literature search with domain expertise to create high-quality labels for *Plasmodium falciparum* specific proteins that distinguish between antigen candidates and intracellular proteins. Additionally, we used our benchmark to compare different well-known prediction models and available protein localization prediction services on the task of identifying protein antigen candidates. We show that available general-purpose services are unable to provide sufficient performance on identifying protein antigen candidates and are outperformed by our models that were trained on this tailored data.
Availability: *PlasmoFAB* is publicly available on Zenodo with DOI 10.5281/zenodo.7433087. Furthermore, all scripts that were used in the creation of *PlasmoFAB* and the training and evaluation of machine learning models are open source and publicly available on GitHub here: https://github.com/msmdev/PlasmoFAB.

II.1 Introduction

Malaria is a major health problem worldwide, causing more than 247 million cases and approximately 619,000 deaths in 2021 [90]. Almost all malaria cases are caused by *Plasmodium falciparum* (*P.falciparum*), predominantly in Africa. Children, pregnant women, and malaria-naïve subjects are at high risk to develop severe malaria [173, 174]. Furthermore, the increase in resistance to both insecticides that target the mosquito vector and anti-malaria drugs, as well as the COVID-19 pandemic, led to an increase of morbidity in several highly endemic countries in the past years [89]. Vaccines are very effective means in protecting against infectious diseases as recently demonstrated in the case of COVID-19. The RTS,S vaccine is the first malaria vaccine recommended by the World Health Organization (WHO) for widespread use in children in endemic settings with a substantial reduction of severe malaria cases, but limited reduction of transmission of malaria[175, 176]. Besides this first success in fighting severe malaria, there is still an urgent need to develop an effective malaria vaccine that confers sterile protection

and reduces malaria transmission. However, developing an effective malaria vaccine is still challenging due to the complex, multi-stage life-cycle of *P.falciparum*, which is genetically highly diverse and employs several immune evasion strategies. As a result, our understanding of immune responses to *P.falciparum*-specific antigens that mediate naturally acquired or experimentally induced protection is incomplete.

More than 5,300 genes are expressed during the life-cycle of *P.falciparum* [177]. However, only a small subset of proteins that are expressed by *P.falciparum* is considered in current target candidate screening processes for an effective malaria vaccine [91, 92]. Since most of the unused proteins have unknown function and experimental validation remains costly and time-intensive, computational methods can be used for pre-screening of proteins of interest. For example, trans-membrane topology prediction is an established task in bioinformatics, where the aim is to predict how and if a protein resides in the cell membrane, i.e., predict the location and length of trans-membrane domains. The class of membrane proteins is one of the most important classes of proteins for medical use. About 25-30% of natural proteins reside in the cell membrane and are, thus, often bound by antibodies during an immune reaction [178]. Another class of relevant proteins for vaccine and drug development is the class of exported proteins. Many of these fulfill important functions for parasite survival. For example, certain proteins ensure that infected red blood cells stick to the microvasculature, one of the factors that makes malaria a potentially fatal disease [179, 180]. In recent years, several scholars developed general-purpose models for sub-cellular localization prediction and offered them as prediction services to be used by the academic community [26, 27, 28, 29, 30]. While general-purpose models provide researchers with an easy-to-use solution for performing prediction tasks, the lack of out-of-distribution generalization capabilities of most general-purpose models leads to sub-optimal prediction performances on novel datasets and misleading pre-screening results [24]. However, training supervised machine learning models for protein antigen candidate prediction needs a sufficient amount of protein sequences with high-quality labels. Currently, only a small fraction of publicly available *P.falciparum* protein sequences have high-quality labels, making the training of models for identification of such antigens for vaccine and drug development exponentially harder. With this work, we introduce the ***Plasmo****dium* ***F****alciparum-specific* ***A****ntigen candidate* ***B****enchmark* (*PlasmoFAB*), a manually pre-processed and curated dataset containing labeled protein sequences for *Plasmodium falciparum* protein antigen candidate prediction.

This manuscript is structured as follows. We describe in detail the process of creating *PlasmoFAB* including the used data sources, pre-processing, and validation steps. Afterwards we present our experiments for predicting *P.falciparum* protein antigen candidates. Here we show the limitations of using established tools and present approaches that provide solutions to overcome these limitations. We conclude our work with a discussion about necessary actions that have to be taken in order to further improve *PlasmoFAB* and, hence, further foster the development of vaccines and drugs to control malaria.

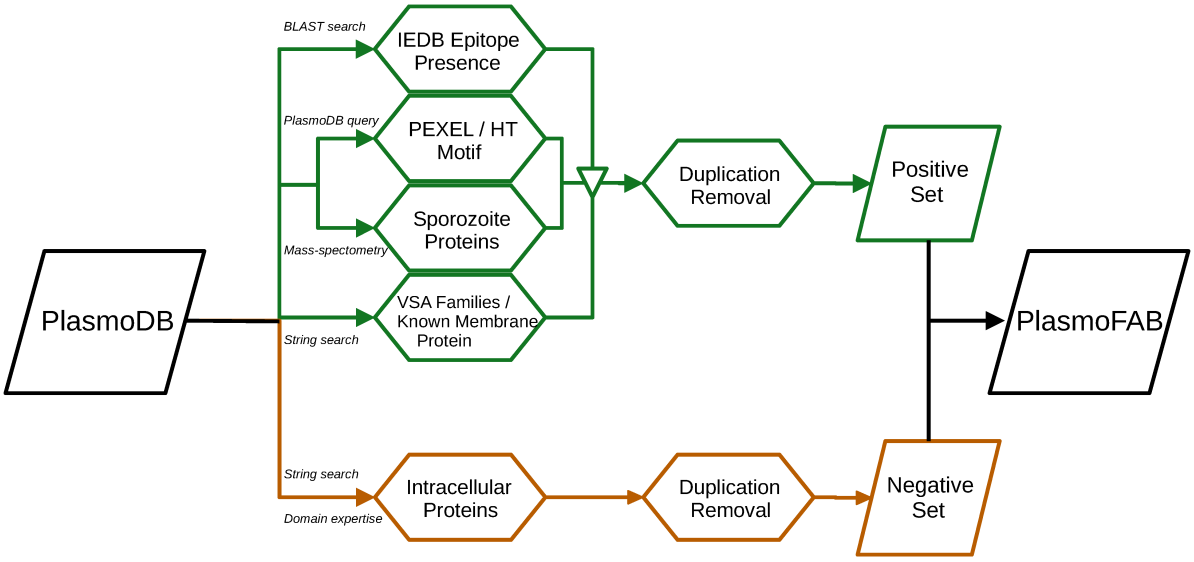

Figure II.1: Schematic overview of the pre-processing steps for the creation of *PlasmoFAB*. The green workflow shows the pre-processing of the positive set, i.e., *P.falciparum* proteins that are either extracellular or membrane-located which renders them eligible to be considered as antigen candidates. We used knowledge-driven techniques like algorithmic homology search, mass-spectrometry, string search, and validation by published literature to create sets of proteins containing antigen candidates. These sets were merged and duplicates were removed to create the positive set. The red workflow shows the pre-processing of the negative set. We combined enzymes validated by published literature with proteins that were assigned to be intracellular by a domain expert and UniProtKB/SwissProt (reviewed) to create the negative set.

II.2 *PlasmoFAB*: Plasmodium Falciparum-specific Protein Antigen Candidate Benchmark

The term supervised machine learning (SL or supervised ML) summarizes techniques that correlate patterns within datasets to desired output variables, i.e., labels for classification or continuous values for regression. The foundation of using supervised ML methods for scientific discovery in medical research are curated datasets with validated and biologically meaningful output variables. Currently, there is no benchmark that fulfills this prerequisite for the exploration of *P.falciparum* protein antigen candidates. With this manuscript, we tackle this fundamental obstacle for supporting *P.falciparum* protein antigen candidate exploration with supervised ML techniques.

In the humoral immune response, the production of antibodies is an important step in getting rid of pathogens. To enable this response chain, pathogen-specific antigens activate B-cells and their differentiation into antibody secreting plasma cells. Therefore, an antigen candidate has to be visible by the humoral immune system of the host. *P.falciparum* protein antigen candidates can be considered visible, if they are present on the outside of infected host cells, like surface proteins, transmembrane proteins, membrane-located proteins or exported proteins. The VEuPathDB database *PlasmoDB* [181] contains the complete genome of different

Table II.1: Composition of the *PlasmoFAB* benchmark. The difference between the sum of sequences in each inclusion criterion and the total number of unique sequences in the positive set occurs due to the fact that some proteins fulfill more than one inclusion criterion. These proteins were not duplicated, resulting in the mismatch between the sum of proteins in each criterion and the total number of proteins in *PlasmoFAB*.

Positive Set (Unique Total = 438)			**Negative Set (Unique Total = 384)**		
Inclusion Criterion	Identified By	# Proteins	Inclusion Criterion	Identified By	# Proteins
IEDB epitope	BLAST match (high confidence)	57	Intracellular Proteins	Combined string and literature search; Domain expertise	384
IEDB epitope	BLAST match (medium confidence)	60			
PEXEL/HT motif	*PlasmoDB* query	265			
Sporozoite proteins	Mass-spectrometry [99]	13			
VSA family / Membraneproteins	Combined string and literature search	302			

Plasmodium species. The protein sequences of the reference strain 3D7 of *P.falciparum*, available in *PlasmoDB*, are the data source for our curated benchmark. We only selected sequences with experimental evidence, i.e., the corresponding *P.falciparum* protein has to be referenced in published work with a unique publication identifier. However, these sequences do not have a sub-cellular location label. We combined an extensive literature search with domain expertise to create high-quality sub-cellular location labels that can be used to train ML models on the task of protein antigen candidate prediction for *P.falciparum*. In other words, *PlasmoFAB*'s positive set contains *P.falciparum* proteins that are accessible at the surface or the exterior of infected cells, like surface proteins, transmembrane proteins, membrane-located proteins, or exported proteins. On the other hand, *PlasmoFAB*'s negative set contains intracellular proteins, which are needed by the parasite to maintain the intracellular life cycle in hepatocytes or erythrocytes. The executed pre-processing steps for the creation of *PlasmoFAB* are detailed in the following section. A schematic overview of our pre-processing can be found in Figure II.1 and the basic statistics of *PlasmoFAB* are shown in Table II.1.

II.2.1 IEDB Epitopes

An epitope is the part of an antigen that is recognized by the immune system of a host organism, i.e., the binding site of an antibody. The Immune Epitope Database (IEDB, https://www.iedb.org/, [182]) contains sequences of known epitopes. We used exact string matching and BLAST similarity matching to compare *P.falciparum* protein sequences with sequences contained in the IEDB. Proteins that either contained exact matches of epitope sequences or a positive BLAST hit with high or medium confidence score were labeled as antigen candidates for our benchmark.

II.2.2 PEXEL/HT Motif

The majority of *P.falciparum* proteins that are either exported into the extracellular space by the parasite or integrated into the membrane of infected erythrocytes contain a specific amino acid sequence called Plasmodium exported element (PEXEL) or host targeting (HT) [183, 184]. Therefore, the presence of this motif is a strong indicator of a protein antigen

candidate. *PlasmoDB* indicates the presence of the PEXEL/HT motif within a sequence by a flag in one of its data fields. For *PlasmoFAB*, we included all proteins with the PEXEL/HT motif as positive antigen candidates.

II.2.3 VSA families and known membrane proteins

Variant surface antigen (VSA) families describe proteins that are typically located on cell surfaces. There are three known VSA families in the *P.falciparum* genome: Plasmodium falciparum erythrocyte membrane protein 1 (PfEMP1), repetitive interspersed family (RIFIN), and sub-telomeric variable open reading frame (STEVOR). The first family, PfEMP1, summarizes proteins that are expressed on the surface of infected erythrocytes during the trophozoite and schizont stage of the infection cycle. These proteins are mainly responsible for effective evasion of immune responses [180]. Proteins belonging to the RIFIN family are exported onto the cell surface of infected erythrocytes as well. They mediate the sequestration of erythrocytes which results in erythrocyte rosetting that further helps parasites to evade immune responses and can block the blood flow [180]. Similar to the other two VSA families, STEVOR proteins are also used by *P.falciparum* parasites to evade host immune responses. They play active roles in the trophozoite, schizont, merozoite, and gametocyte stages of the infection cycle [185, 180]. Beside the members of VSA families, there are a number of known membrane proteins. In the sporozoite stage, those include thrombospondin-related anonymous protein (TRAP), also known as sporozoite surface protein 2 (SSP2), apical membrane antigen 1 (AMA1), liver stage antigen 1 (LSA1), and exported protein 1 (Exp-1), also known as circumsporozoite-related antigen (CRA). Additionally, we included known surface proteins that can be found in other stages of the infection cycle like the family of monomeric serine-threonine protein kinases (FIKK, [186]), the helical intersperse sub-telomeric family of exported proteins (PHIST, [187]), and the multigene family of cytoadherence linked asexual gene (CLAG, [188]).

Each entry in *PlasmoDB* has a textual product description field containing information about the sample in textual form. We performed a string search on the textual product description field using the names of the VSA families as search terms: '*PfEMP1*', '*RIFIN*', '*STEVOR*'. For additional known membrane and exported proteins, we did not only included the names but also descriptive search terms since the textual product description field is not standardized. The additional search terms were '*surface*', '*circumsporozoite*', '*membrane*', '*exported*', '*serine repeat antigen*', '*TRAP*', '*FIKK*', '*GLURP*', '*CLAG*', '*PHIST*', and '*GPI-anchor*'. However, the source and rationale behind the annotation in *PlasmoDB*'s textual product description field are not always disclosed. To ensure that only validated membrane and exported proteins are included in our benchmark, we performed a literature search for each protein that was selected by our string search and included only proteins with published experimental evidence into our benchmark. To further enrich the set of known membrane proteins, we added a list of sequences validated by the UniProtKB/SwissProt (reviewed) database. This database contains high quality, manually annotated proteins sequences [189].

II.2.4 Sporozoite surface-exposed proteins

The authors in [99] used mass-spectrometry to identify potential surface-exposed sporozoite proteins of *P.falciparum*. They assigned priority scores to each investigated protein ranging from 1 (high confidence) to 6 (low confidence). We downloaded the publicly available data from [99] and selected all proteins with a priority score from 1 to 3. We used the unique transcript ID of these proteins to merge this information into the *PlasmoDB* data table and included them into our benchmark as antigen candidates.

II.2.5 Intracellular proteins

The pre-processing steps described above added positive samples, i.e., *P.falciparum* protein antigen candidates, to our benchmark. However, *PlasmoFAB* needs negative samples, i.e., proteins that are not *P.falciparum* protein antigen candidates, to be usable for training of supervised ML methods. A model can only learn to detect true protein antigen candiates, if a set of high-quality negative samples, a so-called negative set, is available. Similar to the positive samples, we curated the negative samples to ensure that only intracellular *P.falciparum* proteins are included into the negative set. Intracellular proteins can only leave the cytoplasm in specific situations that do not reliably occur in the infection cycle, like the burst of an infected erythrocyte or if macrophages digest an infected erythrocyte and subsequently present an intracellular protein as an antigen. However due to the unreliability of these incidents and the fact that both can only occur late in the infection cycle, intracellular proteins are not suitable as antibody targets. Enzymes constitute a subset of intracellular proteins. We performed a string search with the term '**ase*' on *PlasmoDB*'s textual product description field and included all proteins with published experimental evidence of being enzymes into the negative set of our benchmark. While there is a small number of enzymes that are exported to the cell membrane, we made sure to exclude all enzymes from the negative set for which published experimental evidence of being membrane-located exists. Furthermore, we included a list of known intracellular proteins compiled by a domain expert and a list of intracellular proteins validated by UniProtKB/SwissProt (reviewed).

II.3 Utilizing Machine Learning for Plasmodium Falciparum Protein Antigen Candidate Exploration

Manually exploring *P.falciparum* proteins for potential antigen candidates is a time consuming and expensive procedure. With the help of our curated benchmark, we can utilize supervised ML to accelerate the process with a pre-screening of potential proteins that reduces the required workload of researchers in the laboratory. The usefulness of such a pre-screening process highly depends on the accuracy that prediction models are able to achieve. We compared the performance of several ML approaches that are commonly used for textual data, especially for biological sequences. The used methods include a kernelized support vector machine (SVM) utilizing the oligo kernel [53], the protein language model embedding ESM-1b [190] combined with a logistic regression (LR) classifier as well as an SVM, and the protein language model embedding ProtT5 [191], which we also combined with an LR classifier and an SVM. Furthermore, we also tested the performance of existing protein localization prediction

tools on the *P.falciparum* protein antigen candidate prediction task. These tools are publicly offered as a service for protein localization prediction tasks and included TMHMM [192], DeepTMHMM [27], DeepLoc 1.0 [28], DeepLoc 2.0 [29], and Phobius [30]. To ensure a fair comparison between pre-trained prediction services and our self-trained models, we defined a test set that was separated from the training data before model training was performed. We used *MMseqs2* [193, 194] to ensure that each sequence in the test set had at most 30% homology to sequences in the training set, which is the default setting of *MMseqs2*. The test set consists of 60 sequences (30 antigen targets and 30 intracellular proteins) with the remaining 788 sequences in *PlasmoFAB* used as a training set. All performance measures shown in this section are computed on the test set.

To assess the performance of each method, we used three performance measures that are widely used in computational biology due to their ability to handle imbalanced data with relative ease. First, we used balanced accuracy, which has different definitions in literature. We used the arithmetic mean of sensitivity and specificity [15] given by

$$\mathrm{Acc}_{\mathrm{bal}} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right), \tag{II.1}$$

where TP is the number of correctly predicted protein antigen candidates (i.e., true positives), FP is the number of wrongly predicted protein antigen candidates (i.e., false positives), TN is the number of correctly predicted intracellular proteins (i.e., true negatives), and FN is the number of wrongly predicted intracellular proteins (i.e., false negatives). Additionally, we used the F_1-score that is the harmonic mean of precision and recall [16] given by

$$\mathrm{F}_1 = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}, \tag{II.2}$$

with TP, FP, and FN defined in the same way as above. Finally, we also included the Matthews correlation coefficient (MCC, [9]), which is widely recognized as one of the most reliable performance measures for binary classification on biological data. The MCC is defined as

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN}-\mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}. \tag{II.3}$$

Again, the definition of TP, FP, TN, and FN are the same as above. Since the classes in *PlasmoFAB* are balanced, we also report precision, recall, and specificity to provide a quick overview over the distribution of FN and FP for the predictions of the tested models.

II.3.1 Using *PlasmoFAB*'s training sequences for model training

Hyperparameter optimization and model selection was exclusively performed on *PlasmoFAB*'s training sequences to avoid information leakage from the test sequences. As a baseline model, we trained a kernelized SVM utilizing the oligo kernel, a kernel function that was specifically developed for biological sequences [53]. This kernel computes the similarity of two sequences based on k-mer occurrence with a tunable degree of positional uncertainty. The SVM that was trained for *P.falciparum* protein antigen candidate prediction had three hyperparameters that needed to be optimized: the k-mer length, the positional uncertainty parameter σ, and

the regularization parameter C_{SVM}. We performed a grid search utilizing repeated nested cross-validation to optimize all three hyperparameters. The resulting choices were $k = 1$, $\sigma = 18$, and $C_{\mathrm{SVM}} = 0.001$.

Additionally, we used two more complex language embedding models that are commonly used for biological sequences: ESM-1b and ProtT5. The first, ESM-1b, is a pre-trained transformer model [190], which is offered as a feature generator for downstream prediction models. It was developed to be used on biological sequences. ESM-1b follows the self-supervised bidirectional encoder representation from transformation (BERT) pre-training procedure. This language model is a transformer architecture with 33 layers and utilizes self-attention with 20 attention heads. The resulting features have a dimensionality of 1280 with a token context size of 1024. ESM-1b was trained on sequence clusters derived from the UniProt database [195]. We refer the interested reader to the original publication for all technical details about ESM-1b. The token context size together with a positional encoding of fixed length limits input sequences to a maximum of 1024 characters. Since there is a significant number of sequences in *PlasmoFAB* that exceed this character limit, we followed published recommendations to cut the middle part of sequences that exceed the 1024 character limitation [29] to be able to use ESM-1b on our benchmark. In total, 261 sequences were affected by this cutting procedure. The computed feature embeddings were used as inputs for the two tested downstream prediction models, LR and SVM. Again, we exclusively optimized the regularization parameters C_{LR} and C_{SVM}, respectively. After performing the grid search, the optimal parameter choices were $C_{\mathrm{LR}} = 0.15$ and $C_{\mathrm{SVM}} = 20$.

The second language embedding that we used was ProtT5-XL-UniRef50 (ProtT5, [191]). This transformer model, based on the language model T5 [196], is specifically developed for biological data and prediction tasks. Similar to ESM-1b, ProtT5 acts as a feature generator for downstream prediction models. In contrast to other language models, ProtT5 follows an encoder-decoder approach and uses a simplified BERT training objective. The architecture employs 24 layers and also utilizes self-attention with 32 attention heads. ProtT5 has an embedding dimensionality of 1024. Since ProtT5 does not use a positional encoding of fixed length but learns a positional encoding for each attention head, the length of input sequences is not limited in theory. ProtT5 was first pre-trained on the BFD database [197] and fine-tuned on UniRef50 [198]. We refer the interested reader to the original publication for all technical details about ProtT5. Although sequence length is not limited when using ProtT5, finite computation power limits the usable sequence length in practice. With the computing resources available to us, an Nvidia Tesla V100 with 32GB RAM, the maximal usable sequence length was 6000 residues. Longer sequences were shortened in the same way we shortened sequences for ESM-1b. Five sequences in *PlasmoFAB* were affected by this reduction of sequence length. Again, we used the feature embedding as inputs for the two downstream prediction models, LR and SVM, and optimized the regularization parameter via a grid search. The optimal parameters were $C_{\mathrm{LR}} = 0.2$ and $C_{\mathrm{SVM}} = 2.0$.

II.3.2 Evaluating prediction models on *PlasmoFAB*'s test sequences

Table II.2 shows the performance of all models on *PlasmoFAB*'s test set. The models trained by ourselves can be directly applied to the test set. Since the publicly available prediction services

Table II.2: Performance of trained prediction models and prediction services on *PlasmoFAB*'s test set. We trained different models on *PlasmoFAB*'s training set including a support vector machine utilizing the oligo kernel (SVM_{oligo}), a combination of the a linear regression with either ESM1b or ProtT5 language model embedding (LR_{ESM1b} and LR_{ProtT5}), and a support vector machine combined with either ESM1b or ProtT5 language model embedding (SVM_{ESM1b} and SVM_{ProtT5}). Furthermore, we used publicly available, pre-trained prediction services on *PlasmoFAB*'s test set. These services include Phobius, TMHMM, DeepTMHMM, Deeploc 1.0, and Deeploc 2.0.

Model	**MCC**	**F1**	**Bal. Acc.**	**Precision**	**Recall**	**Specificity**
SVM_{oligo}	0.3145	0.5882	0.6500	0.7143	0.5000	0.8000
LR_{ESM1b}	0.7071	0.8000	0.8333	**1.0000**	0.6667	**1.0000**
SVM_{ESM1b}	0.7071	0.8000	0.8333	**1.0000**	0.6667	**1.0000**
LR_{ProtT5}	**0.7338**	**0.8235**	**0.8500**	**1.0000**	0.7000	**1.0000**
SVM_{ProtT5}	0.6917	0.8077	0.8333	0.9545	0.7000	0.9666
DeepTMHMM	0.4395	0.6909	0.7167	0.7600	0.6333	0.8001
DeepLoc 2.0	0.4009	0.7079	0.7009	0.6000	0.6923	0.7095
DeepLoc 1.0	0.2691	0.6071	0.6357	0.5667	0.6538	0.6176
TMHMM	0.3015	0.6316	0.6500	0.6667	0.6000	0.7000
Phobius	0.2722	0.6667	0.6333	0.6111	**0.7333**	0.5333

do not always provide a binary output, we converted the prediction output for each service into a binary label. TMHMM and Phobius provide topology predictions for input sequences and we assigned a positive label to all samples with at least one predicted trans-membrane helix or at least one predicted extracellular region. Otherwise the sample was assigned a negative label. DeepTMHMM refines the prediction of TMHMM by providing a label for each residue in an input sample. For the DeepTMHMM output, we assigned a positive label to all samples with residues that had the membrane domain label ('*M*') assigned. Furthermore, a positive label was assigned to samples where DeepTMHMM predicted the outside cell label ('*O*') for all residues. If none of these conditions was fulfilled, the sample was assigned a negative label. DeepLoc 1.0 and 2.0 are tools for subcellular localization prediction and, hence, offer a multi-label output. Each label corresponds to a different subcellular localization. We used the top predicted label for each input sample. If this label was '*cell membrane*' or '*extracellular*', the sample was assigned a positive label, otherwise a negative label was assigned.

Our results show that models directly trained on *PlasmoFAB* training set clearly outperform the available prediction services. The best performance was achieved by combining ProtT5 feature embedding with logistic regression. None of the tested prediction services was able to achieve a comparable performance to the specialized models.

II.4 Discussion

Computational antigen pre-screening with machine learning methods can drastically reduce time- and resource-consuming experimental exploration procedures and, thereby, accelerate

development of drugs and vaccines. However, these computational pre-screening methods heavily depend on high-quality data to produce reliable results. In this work, we take important steps towards utilizing computational pre-screening for Malaria drug and vaccine development by providing *PlasmoFAB*, a benchmark that consists of *Plasmodium falciparum*-specific protein sequences with curated labels that distinguish between protein antigen candidates and intracellular proteins.

Experimental validation is the gold standard to determine subcellular localization labels for proteins. We ensured that each label in PlasmoFAB achieves this gold standard or, if experimental validation is not feasible, comes as close to the gold standard as possible. As detailed in section II.2, the biggest subgroup of proteins that were assigned as antigen candidates was the group of VSA family members and known membrane proteins. We performed an exhaustive literature search and only included proteins into this subgroup for which published experimental evidence exists. Other subsets with experimentally validated labels are sporozoite proteins and proteins that contain the PEXEL/HT motif. Sporozoite proteins were validated by mass-spectrometry [99]. PEXEL/HT motif occurrence is a property of the protein sequence. This property is experimentally validated since *PlasmoFAB* only includes experimentally validated protein sequences. Furthermore, there is experimental evidence that *P.falciparum* parasites use the PEXEL/HT motif to export proteins [184, 183]. This supports our decision to include PEXEL/HT motif occurrence as an indication of protein antigen candidates. The last remaining subgroup in *PlasmoFAB*'s positive set are proteins with known epitopes. IEDB only includes epitopes that are experimentally validated and we used BLAST to perform similarity matching between IEDB entries and *P.falciparum* protein sequences. Although BLAST does not fulfill the gold standard of experimental validation, it is widely considered as the gold standard for sequence similarity matching. By restricting ourselves to BLAST matches with high or medium confidence, we ensured that the reduction in label quality of proteins in this subgroup is minimized. *PlasmoFAB*'s negative set contains two groups of proteins: enzymes and intracellular proteins. We performed an exhaustive literature search to ensure that all included enzymes have experimental evidence of being intracellular. We excluded enzymes, if there is at least one publication with experimental evidence that suggests that the enzyme is being exported outside the cell. The other subgroup, intracellular proteins, were classified by a domain expert. While this does not fulfill the gold standard of experimental validation, we ensured to minimize the reduction in label quality by using domain expertise.

PlasmoFAB uses data that belongs to the *Plasmodium falciparum* strain 3D7. The genome of this specific strain of the *P.falciparum* parasite was the first to be published by Gardner and colleagues in 2002 [185]. It is still today one of the most important information sources for malaria research [176, 175, 92, 91]. Therefore, we made the decision to concentrate on *P.falciparum* strain 3D7 for the first version of *PlasmoFAB*. For future work, we want to further refine *PlasmoFAB* by deriving high-quality labels for protein sequences of other *P.falciparum* strains in order to incorporate as much information about *P.falciparum* protein antigen candidates as possible into our benchmark.

One potentially surprising result is the sub-optimal performance of publicly available prediction services, like DeepTMHMM or DeepLoc 2.0, even though these services are relatively new and show impressive performance capabilities in their respective manuscripts. Our results do not provide evidence that the published performance capabilities of these models are overly

optimistic or that they should not be used in general. On the contrary, we would like to emphasize that prediction services provide a fast and easy-to-use way for researchers without a strong background in machine learning to utilize prediction models in their research or the possibility to use prediction models even if not enough data for model training is available. However, our results highlight one common problem of general purpose models: their lack of out-of-distribution generalization [24]. Models learn certain aspects of the training data's distribution and allow trained models to achieve high prediction performance of unseen data as long as these data points came from the same distribution. However, if those unseen data points came from a different distribution, there is no guarantee that the model will be able to reliably make predictions on the new data. We see this out-of-distribution generalization issue in the relatively poor performance of the used prediction services. Since the *P.falciparum* proteins are likely to be differently distributed than the proteins used to train the prediction services, these services perform poorly when applied to our test set. This result supports our claim that providing curated datasets with high-quality labels for model training is essential for maximising the potential of computational prediction methods on biological prediction tasks like the pre-screening of *P.falciparum* protein antigen candidates. Therefore, our proposed *PlasmoFAB* benchmark offers a solution to one fundamental obstacle in utilizing computational prediction methods in the development process of drugs and vaccines against malaria.

One goal of developing *PlasmoFAB* was to provide the malaria research community with a tool to utilize machine learning in protein antigen exploration processes. However, the potential target user group of *PlasmoFAB* can only benefit from the data if it fulfils two basic requirements. First, potential users have to be enabled to reliably find, access, and reuse data. And second, potential users have to be able to make an informed decision whether the data is applicable for their specific problem. We tackle the first problem by making *PlasmoFAB* publicly available via Zenodo, which is a platform by researchers for researchers that aims to support open science. By uploading our dataset to Zenodo we ensure that the FAIR principles [39] are taken into account. Additionally, we release *PlasmoFAB* in form of comma-separated values (CSV) files. This file format is universally used in different research communities and should maximize the number of researchers that can use our dataset. Furthermore, we created a datasheet for *PlasmoFAB* as described in [40]. With this datasheet, we provide information about the motivation behind creating *PlasmoFAB*, the creation process, the assumptions made, and applicable use cases. Users who are interested in using *PlasmoFAB* can use the datasheet to make an informed decision about the applicability.

II.5 Conclusion

With this work, we introduce *PlasmoFAB*, a new and carefully curated benchmark for the training of models for *Plasmodium falciparum* protein antigen candidate prediction. The benchmark was created by manually validating extracellular, surface-exposed, and intracellular *P.falciparum* proteins to ensure high-quality labels for every sample in the dataset. Such a curated benchmark is an important prerequisite to incorporate learning models into pre-screening protocols for protein antigen candidates.

We furthermore compared commonly used prediction models with publicly available prediction services on the *P.falciparum* protein antigen candidate prediction task. Our results show the

limitations of existing prediction services, which are vastly outperformed by simpler prediction models that are specifically trained for *P.falciparum* protein antigen candidate prediction.

We are confident that our contribution provides a tool that can be used to help the research community to explore the vast number of *Plasmodium falciparum* proteins with unknown functionality and identify new targets for drugs and vaccines against malaria.

Acknowledgement

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. This research was supported by the German Federal Ministry of Education and Research (BMBF) project 'Training Center Machine Learning, Tübingen' with grant number 01|S17054. This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A.

III Inherently Interpretable Position-Aware Convolutional Motif Kernel Networks for Biological Sequencing Data

Jonas C. Ditz Bernhard Reuter Nico Pfeifer

Abstract

Artificial neural networks show promising performance in detecting correlations within data that are associated with specific outcomes. However, the black-box nature of such models can hinder the knowledge advancement in research fields by obscuring the decision process and preventing scientist to fully conceptualize predicted outcomes. Furthermore, domain experts like healthcare providers need explainable predictions to assess whether a predicted outcome can be trusted in high stakes scenarios and to help them integrating a model into their own routine. Therefore, interpretable models play a crucial role for the incorporation of machine learning into high stakes scenarios like healthcare. In this paper we introduce Convolutional Motif Kernel Networks, a neural network architecture that involves learning a feature representation within a subspace of the reproducing kernel Hilbert space of the position-aware motif kernel function. The resulting model enables to directly interpret and evaluate prediction outcomes by providing a biologically and medically meaningful explanation without the need for additional *post-hoc* analysis. We show that our model is able to robustly learn on small datasets and reaches state-of-the-art performance on relevant healthcare prediction tasks. Our proposed method can be utilized on DNA and protein sequences. Furthermore, we show that the proposed method learns biologically meaningful concepts directly from data using an end-to-end learning scheme.

III.1 Introduction

Biological sequences contain valuable information for a wide variety of biological processes. While this property makes them crucial for advances in related research fields, it also provides the potential to improve diagnosis and treatment decisions in healthcare systems. For this reason, a large amount of machine learning approaches that solve learning tasks on biological sequences were developed over the last years. Among others, these approaches include the prediction of splice sites [199] and translation initiation sites [200], predicting binding affinity between proteins and DNA/RNA [60, 94], drug resistance prediction [93], or the denoising of biological sequence data [201]. However, trained models can only be safely incorporated into medical routines if their prediction outcomes can be thoroughly interpreted and understood even by domain experts, e.g., healthcare providers like medical practitioners, without strong knowledge in the foundations of machine learning. Kernel methods and statistical models provide the possibility to interpret results within the data's domain, hence, allowing domain experts to judge outcomes using their own expertise. Yet, scalability issues in terms of data size limit their utility considering the rapid increase of available data in medical and biological research. On the other hand, gradient-based learning approaches like neural networks can handle huge data pools with relative ease but are normally developed as black-box models. Although there are model-agnostic techniques to interpret these models, e.g., saliency maps [202] or Shapley additive explanations (SHAP) [76], recent work by Rudin [83] advises the

use of inherently interpretable models for high stakes scenarios over *post-hoc* explaining black-box models. One problem of *post-hoc* ML explanation models identified by Rudin is their unfaithfulness regarding the original model's computation, which can result in misleading explanations. Sixt and colleagues showed this unfaithfulness for attribution methods by proving that most methods ignored later layers of a model when computing explanations [84]. Furthermore, Bordt and colleagues showed the limitations of *post-hoc* explanations in adversarial contexts [85]. Lipton warned about the danger of optimizing *post-hoc* methods to produce plausible but misleading explanations [116]. In high stakes scenarios like healthcare, decisions made on misleading or wrong explanations can cause dangerous situations with the potential to further harm patients or other vulnerable groups.

In recent years, several efforts were published to combine kernel functions and neural networks [64, 203, 65, 66]. Combining these two approaches enhances neural networks with the interpretability and robustness of kernel methods. On the other hand, it allows to extend learning within a reproducing kernel Hilbert space (RKHS) to problems with massive numbers of data points. Recently, Chen and colleagues introduced these efforts into data mining on biological sequences by developing convolutional kernel networks based on a continuous relaxation of the mismatch kernel [67]. Although these models show promising performance, the choice of kernel resulted in the necessity of a *post-hoc* model for interpretation. Another limitation results from the fact that the mismatch kernel restricts considered k-mer occurrences to a position-independent representation [48, 47]. In many medical tasks, however, positional and compositional variability provide key information. One kernel network approach that utilizes positional information is the recurrent kernel network (RKN) proposed by Chen and colleagues[68]. Another recent approach to incorporate positional information was proposed by Mialon and colleagues [204]. They utilizes a fixed matrix to introduce positional information. While these architectures showed promising performance capabilities, the chosen architectures resulted once again in black-box models with the need for *post-hoc* interpretation. The oligo kernel proposed by Meinicke and colleagues is able to model positional variability and can additionally provide traditional monomer-based representations as well as position-independent k-mer representations as limiting cases [53]. Furthermore, the oligo kernel allows for intuitive and simple interpretation of k-mer relevance and positional variability. However, the oligo kernel cannot be directly incorporated into a convolutional network architecture and does not take into account information provided by compositional variability of motifs. While k-mers are short sequences with fixed letters at each position, motifs are short sequence patterns that can represent more than one possible letter at each position. The above mentioned limitations motivated our work presented here.

This work is structured in the following way. Section III.2 introduces the position-aware motif kernel function and details how to incorporate the position-aware motif kernel into a convolutional kernel layer and how to interpret a trained CMKN model. Section III.3 provides details regarding the conducted experiments on synthetic and real-world data and the results. Finally, section III.4 provides a discussion of presented prediction and interpretation results and section III.5 completes this work with a conclusion.

In summary, our manuscript provides the following contributions:

- We extend convolutional kernel network models for biological sequences to incorporate

positional information and make them inherently interpretable, which removes the necessity for *post-hoc* explanation models. The new models are called convolutional motif kernel networks (CMKNs).
- This extension is achieved by introducing a new kernel function, called position-aware motif kernel, that quantifies the position dependent similarity of motif occurrences.
- We use one synthetic and two real-world datasets to show how our method can be used as a research tool to gain insight into biological sequence data and how CMKNs can provide local interpretation that can help domain experts, e.g., healthcare providers, to quickly interpret and validate prediction outcomes of a trained CMKN model with their domain expertise.

III.2 Methods

In the following section, we will introduce our new kernel function and show how this kernel can be used to create inherently interpretable kernel networks.

III.2.1 Position-Aware Motif Kernel

We introduce a new kernel function that incorporates the positional uncertainty of the oligo kernel [53] but is defined for arbitrary sequence motifs. Furthermore, our kernel function can be used to construct a convolutional kernel layer as described by Mairal [66]. Our kernel function is based on two main ideas: First, we introduce a mapping of sequence positions onto the unit circle, which allows us to represent the position comparison term by a linear operation followed by a non-linear activation function. Second, we introduce a k-mer comparison term. This extension enables the kernel function to deal with inexact k-mer matching, which capacitates our kernel function to handle arbitrary sequence motifs. We call our new kernel function position-aware motif kernel (PAM).

The first part of our position-aware motif kernel compares sequence positions. In prior work, e.g., Meinicke et al., 2004 [53] or Mialon et al., 2021 [204], a quadratic term is usually employed to measure the similarity of positions. We utilize a linear comparison term instead. First, all positions are mapped onto the upper half of the unit circle to create unit ℓ_2-norm vectors: $\tilde{p} = \left(\left(\cos\left(\frac{p}{|\mathbf{x}|}\pi\right), \sin\left(\frac{p}{|\mathbf{x}|}\pi\right)\right)^T\right.$, where $|\mathbf{x}|$ denotes the length of the corresponding sequence. Due to the position vectors now having unit ℓ_2-norm, the position comparison term can be written as follows: $-\frac{1}{4\sigma}\left\|\tilde{p}-\tilde{q}\right\|_2^2 = \frac{1}{2\sigma}(\tilde{p}^T\tilde{q}-1)$. This allows us to define the following position comparison kernel function over pairs of sequence positions:

$$K_{\text{position}}(p,q) = \exp\left(\frac{\beta}{2\sigma^2}\left(\tilde{p}^T\tilde{q}-1\right)\right), \tag{III.1}$$

where β is a scaling parameter that compensates for the reduced absolute distance between sequence positions due to the introduced mapping and σ is a positional uncertainty parameter similar to the homonymous σ parameter of the oligo kernel.

The second part of our position-aware motif kernel compares sequence motifs. For biological sequences, a motif describes a nucleotide or amino acid pattern of a certain length. Sequence

motifs can be written in form of a normalized position frequency matrix (nPFM), which is a matrix in $\mathbb{R}_+^{|\mathbf{A}|\times k}$ with $|\mathbf{A}|$ being the size of the alphabet over which the motif is created and k being the length of the motif. An nPFM has to fulfill the additional constraint that each column has unit ℓ_2-norm (see supplement for more details). For two motifs ω and ω' of length k given as flattened nPFMs, i.e., the columns are concatenated to convert the matrix into a vector, we define the following motif comparison kernel function:

$$K_{\text{nPFM}}(\omega, \omega') = \exp\left(\alpha\left(\omega^T\omega' - k\right)\right). \tag{III.2}$$

This function will become one if the two motifs match exactly and will approach zero with increasing difference of the two motifs. The parameter α determines how fast the function approaches zero and, hence, specifies the influence of inexact matching motifs.

We define our position-aware motif kernel by forming the product kernel using the functions introduced in Equation III.1 and III.2 and aggregating the kernel evaluation of all motif-position pairs with a sum. In other words, the position-aware motif kernel for pairs of sequences $\mathbf{x}$ and $\mathbf{x}'$ over an alphabet $\mathbf{A}$ is given by:

$$K_{\text{PAM}}(\mathbf{x}, \mathbf{x}') = C\sum_{p=1}^{|\mathbf{x}|}\sum_{q=1}^{|\mathbf{x}'|} K_0((\omega_p, p), (\omega_q, q)) \tag{III.3}$$

with

$$K_0((\omega_p, p), (\omega_q, q)) = K_{\text{nPFM}}(\omega_p, \omega_q)\cdot K_{\text{position}}(p, q) = \exp\left(\alpha\left(\omega_p^T\omega_q - k\right) + \frac{\beta}{2\sigma^2}\left(\tilde{p}^T\tilde{q} - 1\right)\right).$$

Here, $|\mathbf{x}|$ and $|\mathbf{x}'|$ are the lengths of the respective sequences, ω_p is the motif of length k starting at position p in sequence $\mathbf{x}$ represented as a flattened nPFM, and ω_q is defined analogously to ω_p but for sequence $\mathbf{x}'$. The constant $C = \sqrt{\frac{\pi^2\sigma^2}{2\alpha\beta}}$ results from the derivation of the motif kernel matrix elements as the inner product of two sequence representatives $\phi_{\mathbf{x}}, \phi_{\mathbf{x}'}$ in the feature space of all motifs as detailed in the Supplement.

III.2.2 Extracting a Feasible Kernel Layer using Nyström's Method

Mairal and colleagues showed that a variant of the Nyström method [70, 71] can be used to incorporate learning within a reproducing kernel Hilbert space (RKHS) into neural networks [65, 66]. We use the same approach to construct a finite-dimensional subspace of the RKHS $\mathcal{H}$ over motif-position pairs that is implicitly defined by K_0 and incorporate learning within this subspace into a neural network architecture.

Consider a set of n anchor points $z_1, \ldots, z_n$, where each anchor point is a motif-position pair $z_i = (\omega_{z_i}, p_{z_i})$. We define an n-dimensional subspace $\mathcal{E}$ of $\mathcal{H}$ that is spanned by a set of anchor points, i.e.

$$\mathcal{E} = \text{Span}\left(\phi_{z_1}, \ldots, \phi_{z_n}\right), \tag{III.4}$$

where ϕ_{z_i} denotes the projection of each anchor point into the RKHS $\mathcal{H}$. Utilizing the kernel trick, a motif-position pair can be projected onto $\mathcal{E}$ without explicitly calculating the images of the anchor points $\phi_{z_1}, \ldots, \phi_{z_n}$. This natural parametrization is given by [66]

$$\psi((\omega, p)) = K_{ZZ}^{-\frac{1}{2}} K_Z((\omega, p)). \tag{III.5}$$

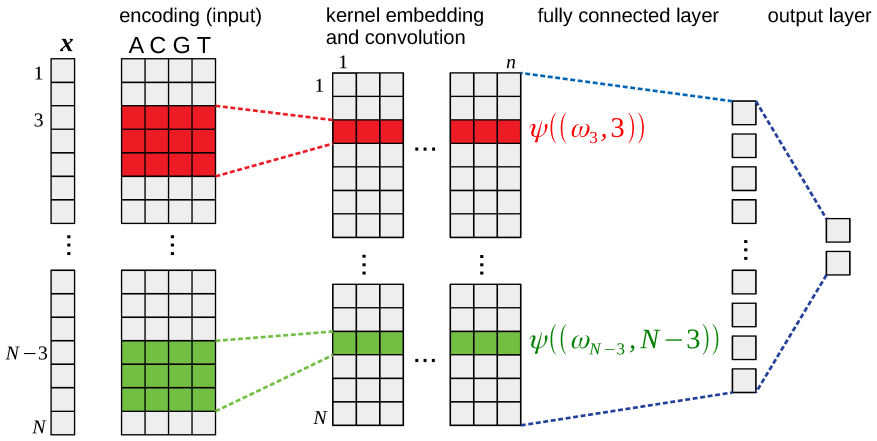

Figure III.1: Schematic overview of an CMKN model. Each motif-position pair of the input is projected onto the subspace of the RKHS by the kernel layer. Afterwards, the projected input is classified using one or several linear fully-connected layers.

Here, $K_{ZZ} = (K_0(z_i, z_j))_{i=1,\ldots,n;j=1,\ldots,n}$ is the Gram matrix formed over the anchor points, $K_{ZZ}^{-\frac{1}{2}}$ is the (pseudo)-inverse square root of the Gram matrix, and $K_Z((\omega, p)) = (K_0(z_1, (\omega, p)), \ldots, K_0(z_n, (\omega, p)))^T$. We follow the procedure proposed in prior work [66, 67] to initialize anchor points. First, we sample a set of $m >> n$ motif-position pairs from the training data. Afterwards, we perform k-means clustering with euclidean distance metric using k-means++ initialization to get n cluster centers of the sampled set. After convergence, we enforce the nPFM constraints onto the cluster centers. With initialized anchor points, CMKN models can be trained by a simple end-to-end learning scheme. A schematic overview over a CMKN model for DNA input together with a visualization of the information flow within the network is shown in Figure III.1.

III.2.3 Interpreting a CMKN Model

The main intuition behind the position-aware motif kernel is to detect similarities between motifs, even if they occur at a certain distance from each other and even if the nPFM underlying the motif is different to a certain degree. In this way, our approach extends previous approaches like the oligo kernel [53] and the weighted degree kernel with shifts [49], which only evaluated exact k-mer matches. However, our kernel is based on a concept that we call motif functions which are extensions of the oligo functions introduced by Meinicke and colleagues [53] (see Supplement for details). A motif function represents the nPFM and position(s) of occurrence of the corresponding motif with a smoothing of the position to account for positional uncertainty. Apart from providing a biologically meaningful feature representation, the use of a kernel based on motif functions allows for a direct interpretation of a trained CMKN model without the need for *post-hoc* methods. If the CMKN model consists only of linear fully-connected layers after the kernel layer, as strictly applied throughout this study, important sequence positions and corresponding motifs can be directly inferred from the learned weights and anchor points, since this ensures that only linear combinations of the learned feature representations are considered. The importance of a sequence position for a certain class can be assessed by

calculating the mean positive weight of the edges that connect the position with the output state that corresponds to the class. The importance ι of position p for class c can thereby be expressed as:

$$\iota_c^p = \frac{1}{|N_p|} \sum_{n \in N_p} \tilde{\iota}_{n,c} \ , \qquad \tilde{\iota}_{n,c} = \begin{cases} \sum_{\{m | m \in N^{(n)}\}} w_{n,m} \tilde{\iota}_{m,c}, & \text{if } N^{(n)} \cap N^{(O)} = \emptyset \\ 1, & \text{if } N^{(n)} \cap N^{(O)} = o_c \\ 0, & \text{otherwise.} \end{cases} \tag{III.6}$$

Here, N_p denotes the set of neurons contributing to the importance of position p, $N^{(n)}$ denotes the set of neurons from the next layer connected by an edge with positive weight to neuron $n \in N_p$, $w_{n,m}$ denotes the weight of the edge connecting neuron n with neuron $m \in N^{(n)}$, and $N^{(O)}$ denotes the set of $|c|$ neurons o_c, each representing a single class, in the output layer. Furthermore, the motif associated with the class at that position is retrieved by identifying all learned motifs with positive weights and calculating the weighted mean motif using the learned weights. This procedure is similar to inferring feature importances from the primal representation using the learned parameters of a SVM. Said utilization of the primal representation is possible for linear kernels and most string kernels [53]. The importance of each amino acid at each position of the motif can be directly accessed by sorting the rows of each column of the associated nPFM in decreasing order. Additionally, motif functions enhance CMKN models with the ability to compute local interpretations, i.e., an explanation of prediction results for single inputs within the data's domain. For an input sequence and a learned motif-position pair, we can estimate the importance of that pair by calculating the ℓ_2-norm of the corresponding motif function. To assess the class that a model associates with an important position, the class-specific motifs that were learned by a model at that position can be retrieved and ranked by the ℓ_2-norm of the motif functions on the input sequence. The motif with the highest ℓ_2-norm determines which class a model assigns to the position. We show an exemplary visualization for domain experts of this procedure in Figure III.3**b**.

III.3 Experiments

We used synthetic data to evaluate CMKN's ability to recover meaningful sequence patterns. Furthermore, we evaluated the performance capability of our proposed method on two different prediction tasks: antiretroviral drug resistance prediction and splice site recognition.

III.3.1 Recovering Meaningful Patterns in Synthetic Data

In order to assess whether CMKN models can reliably recover distinct biological patterns from sequences, we created a synthetic dataset containing 1000 randomly generated DNA sequences of length 100. The set was equally split into negative and positive sequences, with a distinct motif embedded into each class of sequences at a specific position (see Figure III.2**a** for the embedded motifs). For negative sequences, the motif was embedded at position 20 with a positional uncertainty of $\pm$ 5 positions. For positive sequences, the motif was embedded at position 80 with a positional uncertainty of $\pm$ 5 positions. The compositional variablity shown in Figure III.2**a** can be understood in a way that one-third of the 5-mers embedded into negative sequences had a thymine at position 2 while two-third of the k-mers had a cytosine.

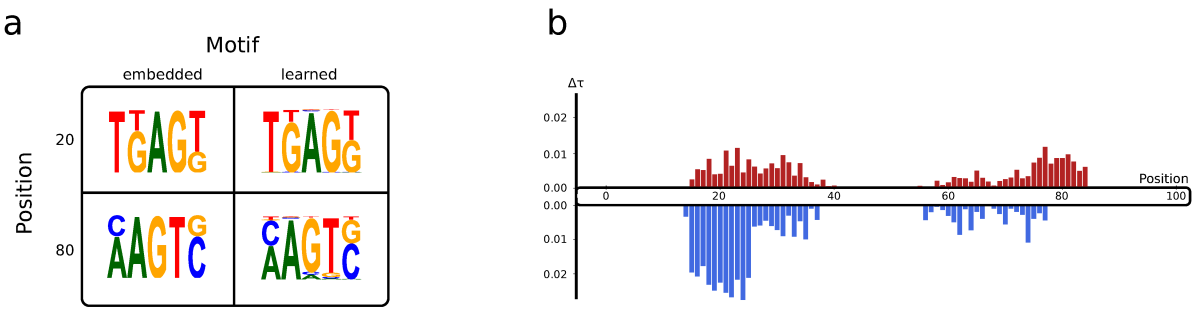

Figure III.2: Evaluation of the interpretation capabilities of CMKN using synthetic data. **a**: The matrix shows the embedded motifs (left column) and the motifs learned by CMKN (right column). The first row shows the motif at position 20 which was only embedded into negative sequences. The second row shows the motif at position 80 which was only embedded into positive sequences. **b**: Positional feature importance of CMKN on the synthetic data. Each bar shows the derivation from the mean positional feature importance for the corresponding sequence position. Red bars indicate importance for the positive class and blue bars indicate importance for the negative class.

This is equivalent for the other motif positions with compositional variability. By creating a synthetic dataset with this procedure, we made sure that the data contains positional and compositional variability that are important for the prediction task. We trained a CMKN model using a motif length of 5 and a positional uncertainty parameter of 4. For the kernel layer, we chose 50 anchor points. The other kernel hyperparameters were set to $\alpha = 1$ and $\beta = 1000$. The model was trained for 50 epochs using the binary cross-entropy with logits loss function.

Figure III.2 shows the results of our experiment with the synthetic dataset. We recovered the positional feature importance values as well as the learned motifs at position 20 and 80 using the procedure described in section III.2.3. As clearly visible on the left side, CMKN is able to recover the two embedded motifs with high similarity using simple end-to-end learning without post-hoc model optimization. Furthermore, the right side of Figure III.2 shows that CMKN is able to detect the relevant areas of biological sequences.

III.3.2 Prediction of antiretroviral drug resistance

When choosing a personalized treatment combination for HIV-infected people, it is crucial to know the resistance profile of the viral variants against available drugs. It has been shown that the genetic sequence of a virus can be used to predict resistance against certain antiretroviral drugs [93]. We performed resistance prediction for drugs representing the three most commonly used antiretroviral drug classes against HIV infections: Nucleoside reverse-transcriptase inhibitors (NRTIs), non-nucleoside reverse-transcriptase inhibitors (NNRTIs), and protease inhibitors (PIs). This prediction task was chosen for evaluation of the proposed method, since it remains an highly important problem in the treatment of HIV infections and the acquired immune deficiency syndrome (AIDS) and is often considered as a role model for precision medicine.

Table III.1: CMKN performance on HIV prediction task

Mean performance and standard derivation of prediction models for three different HIV drug classes: PIs, NRTIs, NNRTIs. Models include polynomial kernel SVMs (SVM_{poly}), oligo kernel SVMs (SVM_{oligo}), random forests (RF), convolutional neural networks (CNN), convolutional kernel networks (CKN_{seq}), and convolutional motif kernel networks (CMKN). Highest values are displayed in bold.

Drug Class	Model	Accuracy	F1 Score	auROC	MCC
PI					
	SVM_{poly}	0.90 ± 0.04	0.83 ± 0.09	0.95 ± 0.03	0.75 ± 0.10
	SVM_{oligo}	**0.92 ± 0.03**	0.86 ± 0.09	**0.97 ± 0.03**	**0.81 ± 0.09**
	RF	**0.92 ± 0.04**	0.85 ± 0.13	**0.97 ± 0.03**	0.79 ± 0.13
	CNN	0.91 ± 0.3	0.84 ± 0.11	0.94 ± 0.05	0.77 ± 0.11
	CKN_{seq}	0.84 ± 0.05	0.72 ± 0.12	0.88 ± 0.05	0.60 ± 0.11
	CMKN	**0.92 ± 0.03**	**0.87 ± 0.09**	0.96 ± 0.03	**0.81 ± 0.10**
NRTI					
	SVM_{poly}	0.86 ± 0.06	0.82 ± 0.09	0.90 ± 0.05	0.70 ± 0.12
	SVM_{oligo}	0.88 ± 0.05	0.85 ± 0.09	**0.94 ± 0.03**	0.75 ± 0.10
	RF	0.88 ± 0.06	0.84 ± 0.12	**0.94 ± 0.04**	0.74 ± 0.15
	CNN	0.88 ± 0.05	0.85 ± 0.09	0.93 ± 0.04	0.74 ± 0.12
	CKN_{seq}	0.79 ± 0.06	0.73 ± 0.12	0.85 ± 0.05	0.54 ± 0.13
	CMKN	**0.89 ± 0.05**	**0.86 ± 0.09**	0.93 ± 0.05	**0.76 ± 0.11**
NNRTI					
	SVM_{poly}	0.82 ± 0.06	0.76 ± 0.11	0.84 ± 0.06	0.63 ± 0.14
	SVM_{oligo}	0.89 ± 0.05	0.86 ± 0.11	0.94 ± 0.05	0.79 ± 0.12
	RF	0.88 ± 0.05	0.85 ± 0.09	0.93 ± 0.07	0.75 ± 0.12
	CNN	0.89 ± 0.04	0.86 ± 0.08	0.94 ± 0.06	0.78 ± 0.10
	CKN_{seq}	0.73 ± 0.06	0.63 ± 0.16	0.78 ± 0.08	0.42 ± 0.15
	CMKN	**0.91 ± 0.03**	**0.89 ± 0.06**	**0.95 ± 0.05**	**0.81 ± 0.08**

Amino acid sequences of virus protein variants with corresponding drug resistance information were extracted from Stanford University's HIV drug resistance database (HIVdb) [205, 206]. An overview of the available data for each of the drugs included in the evaluation can be found in the Supplement. The network architecture used for HIV drug resistance prediction consists of a single convolutional motif kernel layer followed by two fully-connected layers. The first fully-connected layer projected the flattened output of the kernel layer onto 200 nodes and the second fully-connected layer had two output states, one for the susceptible class and one for the resistant class. The motif length and the hyperparameter α of the kernel function were both fixed to 1 based on prior biological knowledge (for details see supplement material). The scaling hyperparameter β was fixed to $\frac{|\mathbf{x}|^2}{10}$ with $|\mathbf{x}| = 99$ for PI datasets and $|\mathbf{x}| = 240$ for NRTI/NNRTI datasets. This compensates for the transformation of sequence positions (for details see supplement material). The number of anchor points and the positional uncertainty parameter σ were optimized using a grid search (for details see supplement material). Due to the limited number of available samples, each model was trained using a 5-fold stratified cross-validation. The data splits for each fold were fixed across models to ensure the same training environment for each hyperparameter combination. Training success was evaluated using the performance measures accuracy, F1 score, and area under the receiver operating characteristic curve (auROC). Due to the fact that some datasets were highly unbalanced, we

also included the Matthew's correlation coefficient (MCC) [207] in the performance assessment.

Mean performances achieved for each of the three investigated drug classes can be found in Table III.1. Our method was able to achieve high accuracy, F1 score, and auROC values for each drug class. Even though the classification problem is highly imbalanced for some of the tested drugs, our model is still able to achieve a high Matthew's correlation coefficient (MCC) value with mean MCC performance exceeding 0.75 for each of the three investigated drug classes. We compared CMKN's performance to previously used models for HIV drug resistance prediction: SVMs with polynomial kernel [93] and random forest (RF) classifiers [208]. Furthermore, we included a SVM utilizing the oligo kernel and the CKN_{seq} model [67] into our analysis. Additionally, we performed an ablation test by replacing the kernel layer with a standard convolutional layer to investigate the influence of our kernel architecture onto prediction performance (denoted by CNN in Table III.1). The results for all models can be found in Table III.1. Our method either outperformed the competitors or achieved similar performance.

III.3.3 Utilizing CMKN's interpretation capabilities to identify resistance mutation positions and motifs

Apart from assessing CMKN's prediction performance, we investigated how well our models were able to learn biologically meaningful patterns from drug resistance data. For each sequence position, we calculated the position importance for each class as described in Section III.2.3 and identified peaks with a sliding window approach, i.e., the mean importance of a window of length 11 around each position was calculated and subtracted from the position importance. We selected the 10 highest peaks identified using this sliding window approach. For each peak position, the associated mean motif (of length one) as well as the two most important amino acids of this mean motif were retrieved using the approach described in Section III.2.3. To get position importance and mean motifs for one of the three investigated drug classes (PIs, NRTIs, and NNRTIs), we averaged the importance values as well as the mean motifs over all models that belong to drugs of the same drug class (8 models for PIs, 6 models for NRTIs, and 3 models for NNRTIs). Figure III.3**a** displays the top ten position of the resistant and susceptible class together with the top two amino acids of the corresponding mean motif for each of the three investigated drug classes. The results indicate that CMKN models are able to learn biologically meaningful patterns from real-world datasets. The most important positions identified by CMKN models correspond mainly to known drug resistance mutation (DRM) positions while the corresponding learned motifs are focused on DRMs. This result is consistent for all three tested drug types. However, CMKN models provide more than a global interpretation. Figure III.3**b** shows the result of CMKN's local interpretation capabilities (as described in section III.2.3) for the model trained on nelfinavir (NFV) data and three randomly selected isolates. First we identified the ten most important positions learned by the model. Afterwards, we retrieved the resistant and susceptible motifs for each position from the trained model. Using the motif functions, we were able to identify which positions the model indicated to be informative for the susceptible class and which positions were indicated to be informative for the resistant class using the procedure described in section III.2.3. This local interpretation shows biologically meaningful patterns and can be used by domain experts to verify a prediction made by the model. For a more detailed discussion of the visualization results, see section III.4.

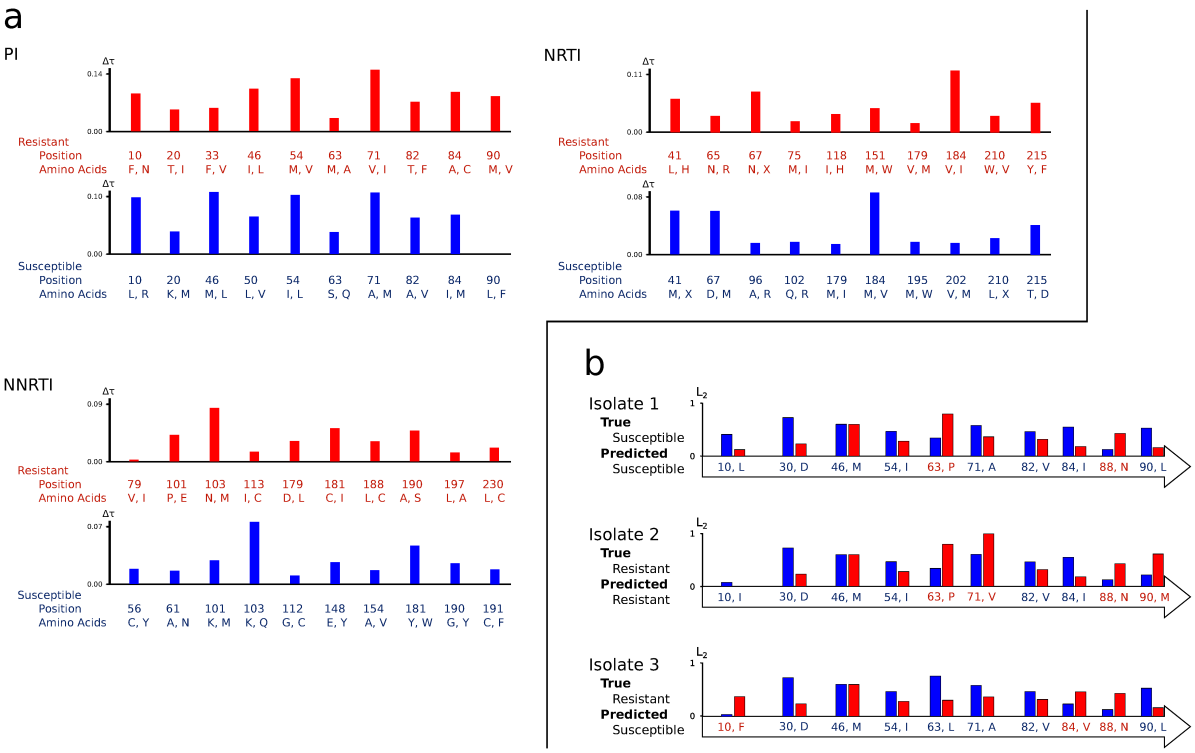

Figure III.3: **a** (Global Interpretation): CMKNs can be used for data mining on biological sequences. The ten most important positions learned by the model, together with the top two contributing amino acids, are displayed. The height of the bar plot at each position indicates the normalized feature importance of that position, i.e., the mean position feature importance was subtracted from the feature importance of the specific position. Higher bars indicate more important positions. The importance of each sequence position was calculated as described in section III.2.3 and peaks were identified using a sliding window approach with a window length of 11. Afterwards, the model's learned motifs associated with the ten highest peaks were calculated (see section III.2.3) and the two amino acids with the highest contribution to these motifs were selected. Positions displayed in red (blue) are associated with the resistant (susceptible) class. **b** (Local Interpretation): We created an exemplary visualization of CMKN's explanation capabilities. Prediction results of the nelfinavir (NFV) model for three randomly chosen input sequences are visualized by showing the learned top ten positions together with the amino acid occurring at the respective position in the input. For each position, the motif functions of the learned motifs are evaluated to identify the one with the highest ℓ_2-norm on the input (see Section III.2.3). If the corresponding motif is a learned resistance (susceptibility) associated motif, the position-amino-acid pair is highlighted in red (blue). The height of the bars above each position corresponds to the ℓ_2-norm of the corresponding susceptible (blue) and resistant (red) motif functions (scaled between 0 and 1). For each isolate, the true and predicted label is displayed.

III.3.4 Splice Site Prediction

The recognition of splice sites is an important task in healthcare, since it can uncover genetic variants and differences in protein composition in individual patients. It consists of two

Table III.2: Test performance on splice site benchmarks. The displayed methods include higher order Markov Chain (MC) classifiers [209], a combination of higher order Markov Chains and SVMs with polynomial kernel (MC-SVM) [210], SVMs with the locality improved kernel (LIK) [209], SVMs with the weighted degree kernel (WD) [209], SVMs with the weighted degree kernel with shifts (WDS) [209], SpliceRover [211], and our CMKN. Highest numbers are shown in bold. Dashes indicate missing values in the original manuscripts.

Model	**NN269**				**DGSplicer**			
	Acceptor		**Donor**		**Acceptor**		**Donor**	
	auROC	**auPRC**	**auROC**	**auPRC**	**auROC**	**auPRC**	**auROC**	**auPRC**
MC	0.97	0.88	0.98	0.92	0.97	0.31	**0.98**	0.42
MC-SVM	0.97	0.88	0.98	0.90	0.95	-	0.95	-
LIK	0.98	0.92	0.98	0.93	-	-	-	-
WD	0.98	0.93	**0.99**	0.93	**0.98**	0.32	**0.98**	0.40
WDS	**0.99**	**0.94**	0.98	0.93	0.97	0.29	0.97	0.36
SpliceRover	**0.99**	-	0.98	-	-	-	-	-
CMKN	0.97	**0.94**	0.98	**0.96**	0.97	**0.65**	**0.98**	**0.65**

classification problems: distinguishing decoys from true targets for acceptor sites and for donor sites.

We used two benchmarks to assess performance of our model on the splice site recognition task: NN269 [212] and DGSplicer [213]. Both benchmarks provide test sets and are highly imbalanced. Details on training and test sets for both benchmarks can be found in the Supplement. For splice site recognition, we used the same architecture that was used for the HIV drug resistance prediction. The hyperparameter α was again fixed to 1. We similarly fixed the scaling parameter to $\beta = \frac{|\mathbf{x}|^2}{10}$ with $|\mathbf{x}| = 90$ for acceptor sequences and $|\mathbf{x}| = 15$ for donor sequences on the NN269 benchmark and $|\mathbf{x}| = 36$ for acceptor sequences and $|\mathbf{x}| = 18$ for donor sequences on the DGSplicer benchmark. The number of anchor points, the motif length k, and the positional uncertainty parameter were optimized using a grid search with 5-fold stratified cross-validation on the training data (details can be found in the Supplement). The model with the best hyperparameter combination was retrained on the whole training set and evaluated using the test set. Training success was evaluated using the area under the precision-recall curve (auPRC), to account for class imbalance, and the auROC to enable comparison with previously published models.

We compared our method to several methods that were previously applied on splice site recognition. These included higher order Markov Chain (MC) classifiers, SVMs with the locality improved kernel (LIK), the weighted degree kernel (WD), and the weighted degree kernel with shifts (WDS) published in [209], a method combining higher order Markov Chains and SVMs with polynomial kernel (MC-SVM) published in [210], and a CNN architecture called SpliceRover [211]. On the NN269 benchmark, our method performed comparable to other methods in terms of auROC and outperformed almost all competitors in terms of auPRC (see Table III.2). On the DGSplicer benchmark, our method performed comparable to other methods in terms of auROC, while substantially outperforming all competitors in terms of auPRC (see Table III.2). An evaluation of CMKN's interpretation on the splice site prediction

task can be found in the Supplement.

III.4 Discussion

In this work, we introduced convolutional motif kernel networks (CMKNs), a convolutional network architecture that allows for end-to-end learning within a subspace of our proposed position-aware motif kernel's RKHS.

By combining a convolutional network architecture with a kernel function, our model is able to perform robust end-to-end learning on relatively small datasets as was shown on data from Standford's HIVdb. Our model was able to generalize to validation data with only a few hundred training samples even in highly unbalanced scenarios. However, due to the fact that our model is based on a standard convolutional network architecture, CMKNs can easily be used on datasets with several hundreds of thousands of samples, as shown on the splice site prediction benchmarks. This allows to utilize our proposed kernel function on very large datasets, something that would be notoriously hard using standard kernel methods like SVMs, since the calculation of a large Gram matrix for our position-aware motif kernel is computationally very demanding.

We included accuracy and auROC as performance measures in our evaluation, since both measures are often used in the ML literature. However, on imbalanced data their informative value is decreased due to a bias towards the majority class [214] as can be seen by considering the auROC vs. auPRC performances on the DGSplicer benchmark in Table III.2. Therefore, we included measures that provide better insights on imbalanced data with few positives: F1 and MCC for HIV drug resistance prediction and auPRC for splice site prediction. Considering F1, MCC, and auPRC, our model performed similar or better compared to all other models.

Another advantage of introducing kernel function evaluation into a neural architecture is the possibility to overcome the black-box nature of neural networks. Since learning within our proposed kernel layer admits a projection onto a subspace of the RKHS of our position-aware motif kernel, each output node of the kernel layer is associated with a position-motif pair. This allows for a biological interpretation of the learned weights associated with each node of the kernel layer. With these global interpretation capabilities, our model can be used as a tool for data mining on biological sequence data. We showed on HIV drug resistance data that our model is able to learn biologically meaningful patterns using standard end-to-end learning methods (see Figure III.3**a**). The majority of the ten most important positions correspond to known DRM positions (nine for PI drugs, eight for NRTI drugs, seven for NNRTI drugs). Furthermore, the top amino acids in the learned resistant motifs reflect known DRMs while the top amino acids in the learned susceptible motifs either reflect the wildtype or none DRMs. There are three exceptions where the susceptible motif features amino acids that lead to an increased drug resistance. These exceptions are leucine (L) and valine (V) at position 50 for PI drugs, valine (V) at position 184 for NRTI drug, and aspartic acid (D) at position 215 for NRTI drugs. However, these exceptions appear to occur due to the averaging of motifs over all drugs for a specific drug class. While all of the four mentioned mutations cause an increase resistance against a subset of drugs [215, 216, 217], they are also a cause of increased susceptibility or have no effect for other drugs [205, 218, 219]. Valine at position 50 reduces susceptibility to

fosamprenavir (FPV), lopinavir (LPV), and darunavir (DRV) but increases susceptibility to tipranavir (TPV). Leucine at position 50 confers high-level resistance to atazanavir (ATV) but increases susceptibility to all other PI drugs. For NRTI drugs, valine at position 184 reduces susceptibility to lamivudine (3TC) but increases susceptibility to zidovudine (AZT), stavudine (d4T), and tenofovir (TDF). At position 215, a mutation to aspartic acid is a so-called thymidine analog mutation that reduces susceptibility to AZT and d4T but has no effect on susceptibility to all other NRTI drugs.

Apart from the data mining capabilities of our proposed CMKN model, the motif functions enrich our model with the capability to provide local interpretations for prediction results within the data's domain. Figure III.3**b** shows an example of the visualization capabilities of our CMKN model using nelfinavir (NFV) data, one of the PI drugs. The figure was created with the following steps. First, the trained NFV model was used to build the susceptible and resistant motifs for each of the ten most informative resistance positions learned for the NFV drug, as described in Section III.2.3 and III.3.3. Afterwards, we assessed for each position if the model relates the position to the susceptible or resistant calss, as described in Section III.2.3. For the first input, which was correctly classified as susceptible, the visualization shows that the model associated susceptible motifs with each of the positions except for position 63 and 88. However, a domain expert can quickly verify that the model falsely classified that the amino acid asparagine (N) at position 88 indicates resistance, since asparagine corresponds to the wildtype and is therefore in accordance with a susceptible isolate. Furthermore, there is no experimental evidence supporting that position 63 is associated with a resistance causing mutation. Using this knowledge, a domain expert can make an educated decision that the prediction is correct. For the correctly classified resistant input, the model associates resistant motifs with positions 63, 71, 90, and, again falsely, with position 88. Since a mutation to methionine (M) at position 90 causes a strong resistance against NFV [215, 220, 221, 222], a domain expert could again directly validate the prediction result. The interpretation capabilities gain importance in case of a wrongly classified input as shown in the bottom part of Figure III.3**b**. Here a domain expert would see that a susceptibility to NFV was predicted while three positions, 10, 84, and 88, are associated with resistant motifs. We again have the previously described, apparently systematic, error at position 88, but a mutation to valine (V) at position 84 causes a moderate resistance against NFV [223]. Additionally, a mutation to phenylalanine (F) at position 10 is known to be associated with reduced in vitro susceptibility to NFV [215, 224]. Thus, the visualization provides the domain expert with all information needed to treat the prediction outcome with the adequate caution. This shows that utilizing the proposed kernel formulation in our model's architecture, together with the proposed motif functions, can provide a visualization of a trained model's output that helps domain experts to validate the predictions.

III.5 Conclusion

Our convolutional motif kernel network architecture provides inherently interpretable end-to-end learning on biological sequence data and achieves state-of-the-art performance on relevant healthcare prediction tasks, namely predicting antiretroviral drug resistance of HIV isolates and distinguishing decoys from real splice sites.

We show that CMKN is able to learn biologically meaningful motif and position patterns on synthetic and real-world datasets. CMKN's global interpretation can foster data mining and knowledge advancement on biological sequence data. On the other hand, CMKN's local interpretation can be utilized by domain experts to judge the validity of a prediction.

Possible future improvements include investigating a combination of different motif kernel layers to combine different motif lengths and extend the architecture to utilize meaningful combinations of motifs. Another improvement that we want to explore in future work is the extension of the kernel formulation to multi-layer networks while securing the interpretation capabilities.

Acknowledgement

The authors would like to thank Prof. Chung-Chin Lu for providing the DGSplicer benchmark. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. This research was supported by the German Federal Ministry of Education and Research (BMBF) project 'Training Center Machine Learning, Tübingen' with grant number 01|S17054. This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. This work was supported by

IV COmic: Convolutional Kernel Networks for Interpretable End-to-End Learning on (Multi-)Omics Data

Jonas C. Ditz **Bernhard Reuter** **Nico Pfeifer**

Abstract

Motivation: The size of available omics datasets is steadily increasing with technological advancement in recent years. While this increase in sample size can be used to improve the performance of relevant prediction tasks in healthcare, models that are optimized for large datasets usually operate as black boxes. In high-stakes scenarios, like healthcare, using a black-box model poses safety and security issues. Without an explanation about molecular factors and phenotypes that affected the prediction, healthcare providers are left with no choice but to blindly trust the models. We propose a new type of artificial neural network, named Convolutional Omics Kernel Network (COmic). By combining convolutional kernel networks with pathway-induced kernels, our method enables robust and interpretable end-to-end learning on omics datasets ranging in size from a few hundred to several hundreds of thousands of samples. Furthermore, COmic can be easily adapted to utilize multiomics data.
Results: We evaluated the performance capabilities of COmic on six different breast cancer cohorts. Additionally, we trained COmic models on multiomics data using the METABRIC cohort. Our models performed either better or similar to competitors on both tasks. We show how the use of pathway-induced Laplacian kernels opens the black-box nature of neural networks and results in intrinsically interpretable models that eliminate the need for post hoc explanation models.
Availability: Datasets, labels, and pathway-induced graph Laplacians used for the single-omics tasks can be downloaded here. While datasets and graph Laplacians for the METABRIC cohort can be downloaded from the above mentioned repository, the labels have to be downloaded from cBioPortal. COmic source code as well as all scripts necessary to reproduce the experiments and analysis are publicly available at https://github.com/jditz/comics.

IV.1 Introduction

n recent years, artificial neural networks (ANNs) show promising performance when employed to learn correlations between data points and outcome variables. They combine feature extraction and prediction training in a single end-to-end learning scheme lowering the necessary amount of labor put into feature engineering and can be used on very large datasets with relative ease. With the advent of big data and high-throughput data generation techniques in computational biology and healthcare, resulting in an increased number of data points available for the training of prediction models, the use of ANNs in these fields has vastly increased. In computational biology, ANNs showed promising performance capabilities when applied to prediction tasks in regulatory genomics [60, 225, 226] and in biological image analysis [227, 228, 229]. Furthermore, several authors showed the potential of ANNs in healthcare scenarios such as diagnosis [230, 231], drug discovery [232, 233], epidemiology [234], personalized medicine [235], and operational efficiency [236]. However, utilizing ANNs for prediction tasks usually comes with two shortcomings: First, a large amount of data is needed to robustly

train a deep neural network and, second, neural network models operate as black-boxes. While the first problem can be tackled by shrinking the complexity of the neural network, which increases the stability of the model but often leads to a decrease in performance, the second shortcoming is most often addressed using *post-hoc* interpretation models. This technique involves solving a secondary task that utilizes a pre-trained prediction model such that the computed solution provides a humanly understandable interpretation for the results computed by the prediction model. Commonly used methods include Shapley additive explanation (SHAP, [76]), counterfactual explanation using generative models [77], and saliency methods like Layer-wise Relevance Propagation (LRP, [78]), Deep Taylor Decomposition (DTD, [79]), GuidedBP [80], or DeepLIFT [81]. Using *post-hoc* interpretation methods can provide additional information to improve understanding and advance scientific knowledge in low-risk scenarios but they have several properties that render their use in high-risk scenarios potentially problematic. Most *post-hoc* interpretation methods are unfaithful to the computations of the original model [83]. Furthermore, many saliency methods ignore information provided by deeper layers of ANNs [84]. Recent work showed that *post-hoc* methods are limited in adversarial contexts [85] and can be exploited to provide seemingly plausible but misleading explanations [116]. In healthcare, decisions that are made based on wrong or misleading explanations have the potential to cause harm to patients.

Kernel methods can provide both robustness on small datasets and interpretation capabilities within the domain of the data. These methods utilize the kernel trick to solve a prediction task by implicitly projecting data into the reproducing kernel Hilbert space (RKHS) of a kernel function and solve the classification or regression problem within the RKHS. While the use of a kernel functions does not always guarantee interpretation capabilities, there are several kernel functions for biological data that result in interpretable models, e.g., the oligo kernel for sequences [53] or the pathway-induced kernel for omics data [55]. Combining kernel functions with ANNs is a promising direction to increase the robustness of ANN models on small datasets and several efforts in that direction have been published in recent years [64, 203, 65, 66]. Chen and colleagues showed the feasibility of kernel networks for biological sequences by using a relaxation of the mismatch kernel [48] to build convolutional and recurrent neural network architectures [67, 68]. Furthermore, they showed how to use convolutional kernel neural networks on graph-structured data like protein structures [69]. While these models showed promising results and increased robustness, the choice of the kernel function resulted in models that are not intrinsically interpretable. However, we recently showed that a carefully chosen kernel function results in intrinsically interpretable kernel networks for biological sequence data [237]. With this work we introduce Convolutional Omics Kernel Networks (COmic), a neural network architecure that allows for intrinsically interpretable end-to-end learning on (multi-)omics data. This is achieved by using a kernel function based on graph Laplacians of biological networks to project input samples into a subspace of the corresponding reproducing kernel Hilbert space (RKHS) with a variant of the Nyström method. Using max-pooling combined with strictly linear layers for classification results in COmic models that provide global interpretation, while attention layers can be used to create COmic models that provide local interpretation. In this manuscript, we use the definition most commonly found in the interpretable ML literature for global and local interpretation [73]. In simple words, *global interpretation* can be used to answer the question *"How does the trained model make predictions?"*, while *local interpretation* can be used to answer the question *"Why did the*

model make a certain prediction for a specific input?".

We show the performance and interpretation capabilities of COmic models on six different breast cancer microarray cohorts. These cohorts contain microarray gene expression data from patients with breast cancer and were stratified on the occurrence of a relapse within 5 years. We compare our proposed method to 15 previously published approaches including several methods based on support vector machines (SVMs) like network-based SVMs [238], recursive feature elimination SVMs [239], and graph diffusion kernels for SVMs [240, 241] as well as classification by average pathway expression [242], classification by significant hub genes [243], classification by pathway activity [244], and pathway-induced multiple kernel learning (PIMKL, [55]). We show how the projection into a subspace of the RKHS of pathway-induced kernels in combination with linear and attention layers leads to global and local interpretations, respectively. Furthermore, we use the METABRIC cohort [245] to show how COmic models can be used on multi-omics data. On the METABRIC breast cancer cohort, we predicted disease-free survival using gene expression (mRNA) and copy number alteration (CNA) data.

This work is structured as follows. We first introduce COmic by describing the pathway-induced kernel and define the necessary network architecture to build a COmic model. Afterwards, we show how to achieve a globally interpretable COmic model using strictly linear layers and a locally interpretable COmic model using attention layers. We evaluate COmic models on six breast cancer cohorts and show how COmic models can be utilized for multi-omics data. With this manuscript, we introduce a new kernel network architecture that can be both robustly trained on small-scale (multi-)omics datasets and easily utilized for prediction tasks on (multi-)omics datasets with several hundreds of thousands of data points. Furthermore, our method results in intrinsically interpretable models offering global and local interpretations of prediction results.

IV.2 Convolutional Omics Kernel Networks

In the following section, we describe the theoretical background of convolutional kernel networks for prediction tasks on omics-based datasets.

IV.2.1 Pathway-Induced Kernel Functions

The foundation of pathway-induced kernel functions are so-called graph Laplacian matrices. To define these matrices we first assume $G = (V, E)$ to be an undirected graph with vertices $V = \{v_1, \ldots, v_n\}$ and edges E. Furthermore, G is assumed to be a weighted graph with weight matrix $W \in \mathbb{R}^{n \times n}$, where $w_{ij} = w_{ji} \geq 0$ describes the weight of the edge between vertices v_i and v_j. The degree of each vertex $v_i \in V$ is defined as $d_i = \sum_{j=1}^{n} w_{ij}$. The diagonal matrix with the degrees $d_1, \cdots, d_n$ on the diagonal is called the degree matrix D. The unnormalized graph Laplacian $L \in \mathbb{R}^{n \times n}$ is defined as [54]:

$$L \coloneqq D - W \tag{IV.1}$$

Since an unnormalized graph Laplacian has undesirable mathematical properties in case of very broadly distributed degrees within G [54], we use a normalized graph Laplacian instead,

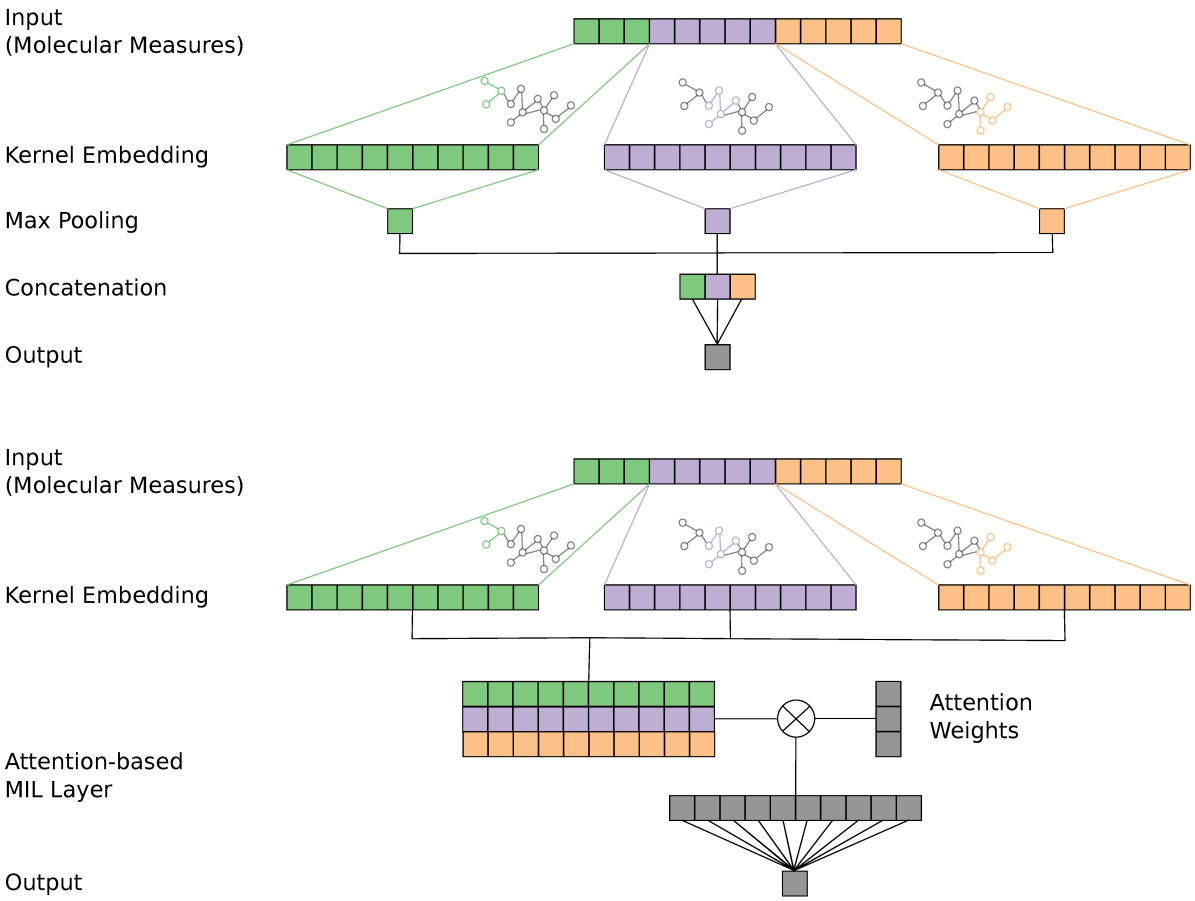

Figure IV.1: Schematic of the proposed interpretable COmic models. **Top:** Pooling-based COmic model. The kernel embedding of each involved pathway are reduced to a single dimension by using a one-dimensional max pooling operation. The output of each pooling layer is concatenated and prediction is performed using a strictly linear fully-connected layer. The pooling-based COmic models are globally interpretable similar to PIMKL models by utilizing the weights of the fully-connected layer. **Bottom:** Attention-based COmic model. The kernel embeddings of each involved pathway are transformed into a bag of instances of a multiple instance learning problem. Attention weights for each instance are calculated using each pathway's kernel embedding and a matrix multiplication between the bag of instances and the attention weights is performed. The output is used for prediction with a strictly linear fully-connected layer. The attention-based COmic models can be locally interpreted by utilizing the attention weights.

which is defined as:

$$L_{\text{sym}} \coloneqq D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \tag{IV.2}$$

Similar to previous manuscripts, see e.g., [246, 55], we use molecular interaction networks (MIN) as graphs underlying the normalized graph Laplacians. Using known interaction networks allows to define a kernel function that computes the similarity of molecular measures (gene expression, DNA methylation, etc.) under the assumed interactions defined by the network. Given two molecular measures $x_i \in \mathbb{R}^n$ and $x_j \in \mathbb{R}^n$, we define the kernel function as

$$K_{\text{MIN}}(x_i, x_j) = x_i^T L_{\text{MIN}} x_j, \tag{IV.3}$$

where L_{MIN} is the normalized graph Laplacian (as defined in Eq. IV.2) of a molecular interaction network.

Manica and colleagues proposed to use pathway-specific sub-networks instead of whole interaction networks for computing normalized graph Laplacians [55]. This method allows for a more tailored induction of prior knowledge into a prediction task. The authors call this approach pathway-induced (PI) kernel functions. Here, the similarity between two molecular measures is not computed using a single graph Laplacian but with a set of p graph Laplacians $\mathbf{L} = \{L_{\text{PI}_1}, \ldots, L_{\text{PI}_p}\}$ each defined over a pathway-specific sub-network of the molecular interaction network. Therefore the pathway-induced kernel for two molecular measures $x_i \in \mathbb{R}^n$ and $x_j \in \mathbb{R}^n$ is not a single function but a set of functions defined as

$$K_{\text{PI}}(x_i, x_j) = \{K_{\text{PI}_1}(x_{i,\text{PI}_1}, x_{j,\text{PI}_1}), \ldots, K_{\text{PI}_p}(x_{i,\text{PI}_p}, x_{j,\text{PI}_p})\} \tag{IV.4}$$

with

$$K_{\text{PI}_r}(x_{i,\text{PI}_r}, x_{j,\text{PI}_r}) = x_{i,\text{PI}_r}^T L_{\text{PI}_r} x_{j,\text{PI}_r}, \tag{IV.5}$$

where $L_{\text{PI}_r} \in \mathbb{R}^{d \times d}$ is the symmetric graph Laplacian of the rth pathway-specific sub-network with d nodes and $x_{i,\text{PI}_r} \in \mathbb{R}^d$ and $x_{j,\text{PI}_r} \in \mathbb{R}^d$ are vectors containing only the signal values of the molecular measures that correspond to nodes within the pathway-specific sub-network described by L_{PI_r}. Manica and colleagues used multiple kernel learning (MKL) to combine the set of pathway-induced kernel functions into a single learning framework. In contrast to prior work, we are using a variant of the Nyström method to formulate an explicit parametrization of an orthogonal projection onto a finite-dimensional subspace of a pathway-induced kernel's RKHS. This enables us to define pathway-induced kernel layers that can be incorporated into artificial neural networks. By using a neural network architecture as the basis for COmic, the resulting models can be tailored to specific datasets (single- or multi-omics) as well as the desired form of interpretation. We will show that in the following sections.

IV.2.2 Convolutional Kernel Layer projects onto a finite-dimensional RKHS-Subspace

Convolutional kernel networks make use of a variant of the Nyström method to project input samples into a finite-dimensional subspace of the RKHS $\mathcal{H}$ of a kernel function. To achieve this, a set of q anchor points $z_1, \ldots, z_q$ is used to define a q-dimensional subspace $\mathcal{E}$ of $\mathcal{H}$. The anchor points lie in the input space of the kernel function and the RKHS subspace is defined as

$$\mathcal{E} = \text{Span}(\phi_{z_1}, \ldots, \phi_{z_q}), \tag{IV.6}$$

where $\phi(z_i)$ denotes the image of the ith anchor point under the kernel function. The orthogonal projection of input points onto $\mathcal{E}$ admits an explicit parametrization that utilizes the kernel trick to avoid explicitly calculating the images $\phi(z_i)$ [66, 70, 71]. For an input x, i.e., a molecular measure in case of omics data, the explicit parametrization $\psi(x) \in \mathbb{R}^q$ is defined as

$$\psi(x) = K_{ZZ}^{-\frac{1}{2}} K_Z(x), \tag{IV.7}$$

where $K_{ZZ} = (K(z_i, z_j))_{i=1,\dots,q;j=1,\dots,q}$ is the gram matrix formed by the anchor points, $K_{ZZ}^{-\frac{1}{2}}$ denotes the (pseudo-)inverse square root of the Gram matrix, and $K_Z(x) = [K(x, z_1), \dots, K(x, z_p)]^T$. As shown in Figure IV.1, each pathway-induced kernel function has to be modelled with a separate orthogonal projection. This means that a COmic model utilizing p pathway-induced kernel functions maps each input onto p representations $\psi_{\mathrm{PI}_1}, \dots, \psi_{\mathrm{PI}_p} \in \mathbb{R}^q$. These representations are then used to solve the prediction task for the input. In the next section, we show two different approaches to combine the representations leading to globally or locally interpretable models, respectively.

The anchor points can be initialized using k-means on all input samples with the number of clusters set to the number of anchor points. Afterwards, the anchor points are optimized with the end-to-end learning scheme used to train the whole network. For all experiments described in this manuscript, anchor points were initialized using k-means++ [247].

IV.2.3 Globally and Locally Interpretable COmic Models

Globally interpretable COmic models are based on multi-kernel learning (MKL). A simple approach to MKL is finding an optimal linear combination of all utilized kernels. This approach learns a weight for each kernel and, therefore, can be used to determine the influence each kernel has on the prediction outcome. Since kernels are directly associated with pathways in PIMKL, Manica and colleagues show that MKL weights can be used to determine the importance of different pathways for a prediction [55]. We can embed a similar weighted sum of pathway-induced kernels into the architecture of COmic models. Each kernel embedding produced by the PI-kernel layer described in section IV.2.2 is passed into a one-dimensional max pooling layer. This results in a single activation $A_r = \max(\psi_{PI_r})$ for each pathway-induced kernel, where $\max(\psi_{PI_r})$ denotes the maximum value in vector ψ_{PI_r}. This activation is high, if the input is similar to one of the learned anchor points, and low otherwise. By concatenating all activations and passing them into a strictly linear fully-connected layer, the model learns a single weight for each pathway-induced kernel and the prediction is calculated as a weighted sum of all kernels, i.e.,

$$\hat{y} = \sum_{r=1}^{p} w_r A_r, \tag{IV.8}$$

where $\hat{y}$ is the overall prediction, $w_r \in \mathbb{R}$ is the weight and $A_r \in \mathbb{R}$ is the activation of the rth pathway-induced kernel. We call this architecture *pooling-based COmic model*. In contrast to the MKL approach, the weights can become negative. This enhances the interpretation capabilities of pooling-based COmic models, since we cannot only infer if a pathway is important for the prediction task but also with which class each pathway is associated by looking at the sign of the weight. The top part of Figure IV.1 shows a schematic of a pooling-based COmic model.

Locally interpretable COmic models are based on multiple instance learning (MIL). In MIL, each sample is represented as a bag of instances with a single label per bag [248, 249, 250]. There are two general approaches to solve an MIL problem: the instance-level approach and the embedding-level approach. In the instance-level approach, an instance-level classifier predicts a score for each of the instances in the bag. Afterwards, scores are aggregated by MIL pooling to compute the prediction for the bag. In the embedding-level approach, a low-dimensional embedding of each instance is computed and MIL pooling is used on the embedded instances to create a bag representation. This representation is used by a bag-level classifier to provide the prediction. While it was shown that the embedding-level approach leads to better performances [251], the instance-level approach leads to interpretable models [252]. Ilse and colleagues proposed an MIL-model based on neural networks that combines the strength of both approaches, called attention-based multiple instance learning [253]. Their approach can be utilized for COmic models in the following way. The output of our proposed PI-kernel layer can be viewed as a bag of low-dimensional instances $H = \{\psi_{PI_1}, \ldots, \psi_{PI_p}\}$, where each $\psi_{PI_r} \in \mathbb{R}^q$ is the projection onto a q-dimensional subspace of the RKHS of one pathway-induced kernel. Attention-based MIL pooling is then used to compute the bag representation, i.e.,

$$\tilde{\psi} = \sum_{r=1}^{p} a_r \psi_{PI_r}, \tag{IV.9}$$

where

$$a_r = \frac{\exp\{w^T \tanh(V\psi_{PI_r}^T)\}}{\sum_{j=1}^{p} \exp\{w^T \tanh(V\psi_{PI_j}^T)\}}. \tag{IV.10}$$

$w \in \mathbb{R}^{l\times 1}$ and $V \in \mathbb{R}^{l\times m}$ are parameters of the attention layer. As noticed by Ilse and colleagues, the $\tanh(\cdot)$ non-linearity introduces a potential limitation due to the fact that it is roughly linear only for $x \in [-1, 1]$. This limitation can be reduced by using a gating mechanism [254]. In this case, the attention weights are calculated as

$$a_r = \frac{\exp\{w^T(\tanh(V\psi_{PI_r}^T) \odot \mathrm{sigm}(U\psi_{PI_r}^T))\}}{\sum_{j=1}^{p} \exp\{w^T(\tanh(V\psi_{PI_j}^T) \odot \mathrm{sigm}(U\psi_{PI_j}^T)\}}. \tag{IV.11}$$

Again, $w \in \mathbb{R}^{l\times 1}$, $V \in \mathbb{R}^{l\times m}$, and $U \in \mathbb{R}^{l\times m}$ are parameters of the attention layer. In both cases, the training of all attention layer parameters is part of the end-to-end training routine for the whole network and, hence, does not introduce the need for additional measures. We call this architecture *attention-based COmic model*. Since the attention weights a_p are input specific, they enhance a model with local interpretation capabilities. The bottom part of Figure IV.1 shows a schematic of an attention-based COmic model.

IV.3 Experiments on Cancer Benchmark Data

To assess the performance and interpretation capabilities of our COmic models we use publicly available cancer benchmarks. The evaluation involves tasks on single-omics data as well as multi-omics data.

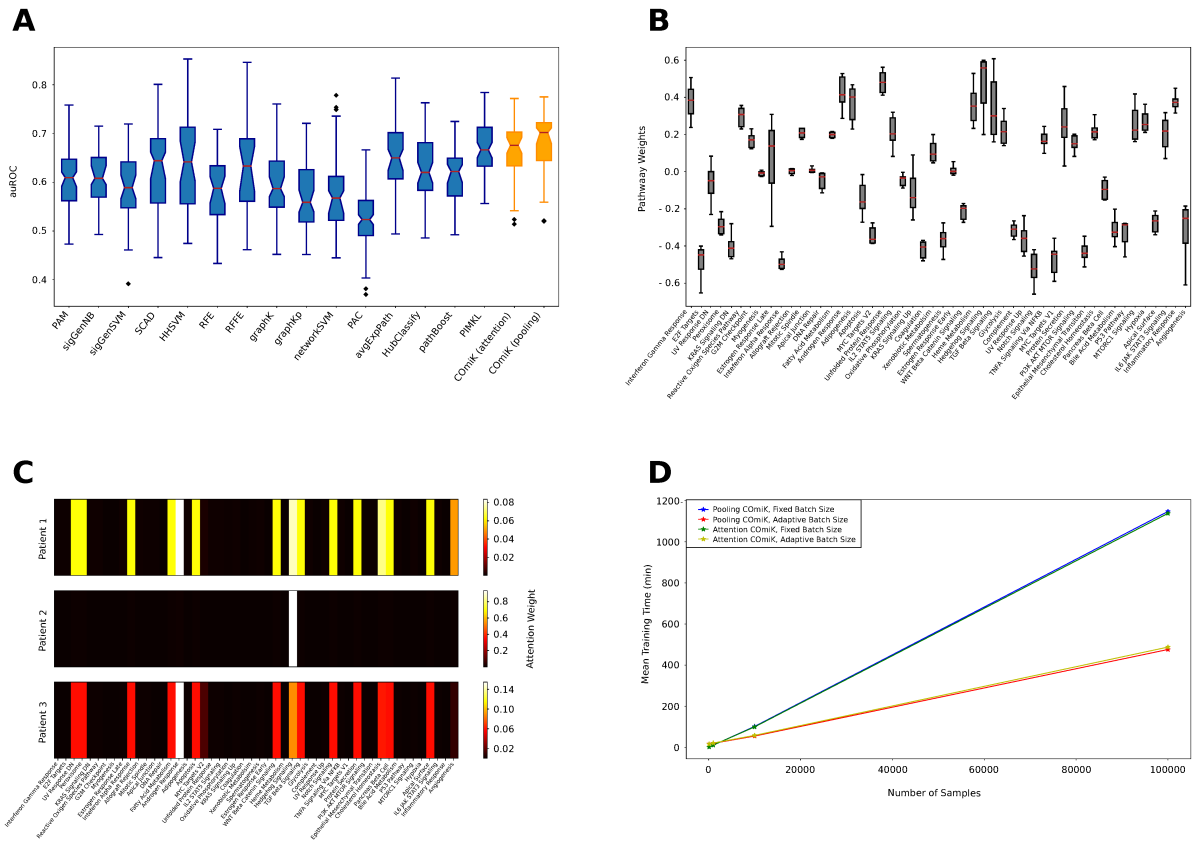

Figure IV.2: **A**: Cross-validation performance of COmic models compared to previously published methods. The boxplots show the ten mean auROC validation scores of a 10-times repeated 10-fold cross-validation over each of the six breast cancer cohorts. Performance of COmic models are shown in orange. The center line of each box indicates the median. The height of the boxes represents the inter quartile range (IQR) with the upper and lower whiskers set to 1.5 times the IQR. Outliers are depicted by black diamonds. Notches represent the confidence interval (CI) around the median and were calculated using bootstrapping with 10000 iterations. **B:** Visualizing the global interpretation capabilities of a pooling-based COmic model. Each box represents one of the 50 pathways and was created using the pathway weights of the models trained on the six publicly available breast cancer cohorts: GSE11121, GSE1456, GSE2034, GSE2990, GSE4922, and GSE7390. The boxplots are defined as in (**A**) but without notches (CIs not shown). **C:** Visualizing the local interpretation capabilities of an attention-based COmic model. Each heat-map shows the attention weights for each of the 50 pathways for three different patients. The model was trained on the GSE11121 cohort. Patient 1 was correctly classified to have a metastasis free survival (DMFS) above five years. Patient 2 was correctly classified to have a DMFS below five years. Patient 3 was wrongly classified to have a DMFS above five years while the DMFS of patient 3 was actually below five years. More examples can be found in the supplement. **D:** Mean training time of pooling-based and attention-based COmic models for differently sized datasets. The number of samples is 100, 1000, 10000, and 100000, respectively. Training was repeated five times per dataset and the stars represent the mean training time. The blue and green lines show the results for a fixed batch size of 32 samples per batch. The red and yellow lines show the results for an adaptive batch size of 1% of the dataset size (i.e., the batch size was 1 for the dataset with 100 samples and 1000 for the dataset with 100000 samples). Each model was trained for 200 epochs.

Table IV.1: Total number of patients, number of patients in each class, and sources of the datasets used in single- and multi-omics prediction experiments. The first six rows contain information about the single-omics datasets used to train COmic models and compare the results to previously published methods. The last line contains information about the METABRIC multi-omics dataset used in multi-omics prediction experiments. The single-omics classes are DMFS/RFS below 5 years and above 5 years while the multi-omics classes are RFS NO and RFS YES.

Dataset	Patients	DMFS/RFS		Source
		< 5y / NO	≥ 5y / YES	
GSE11121	181	28	153	[255]
GSE1456	153	34	119	[256]
GSE2034	275	93	182	[257]
GSE2990	158	42	116	[258]
GSE4922	228	69	159	[259]
GSE7390	191	56	135	[260]
METABRIC	1980	803	1177	[245]

IV.3.1 Single-Omics Prediction on Breast Cancer Benchmark Cohorts

We trained COmic models on six different public breast cancer Affymetrix HGU133A microarray datasets (GSE11121, GSE1456, GSE2034, GSE2990, GSE4922, and GSE7390) that were previously used to benchmark knowledge-based classification methods that use interaction network priors. The task was to predict for each patient if metastasis free survival (DMFS) or relapse free survival (RFS) exceeded five years. On GSE11121 and GSE4922, the end point was DMFS while RFS was considered for all other cohorts. Details about the datasets can be found in Table IV.1. Both, pooling-based and attention-based COmic models, used 50 different pathways to build kernel layers with 30 anchor points each. We used the Laplacians derived from a merge between KEGG pathways and Pathway Commons that were publicly released by Manica and colleagues ([55], see original manuscript and corresponding supplementary material for details). Furthermore, we used gated attention together with an attention dimension of 128 for the attention-based COmic models. Networks were trained for 200 epochs with the Adam optimizer [261] using the class-balanced loss function [262]. The batch size was set to 32. All models presented in this work were trained on a single NVIDIA GeForce GTX 1080 Ti. We used the area under the receiver operating characteristic (auROC) as our performance measure to be comparable to previously published results on the benchmarks. Competitors' performances shown in Figure IV.2A are taken from [55], for the PIMKL model, and [263], for all other competitors.

As shown in Figure IV.2A, COmic models either outperformed competitors or performed similar to previously published methods. Notably, the globally interpretable pooling-based COmic models were able to achieve a small improvement in terms of auROC compared to all other models. On the other hand, the locally interpretable attention-based models achieved a similar performance as the previously best-performing model, PIMKL. We derived exemplary

visualizations to evaluate the interpretation capabilities of COmic models. Since the pooling-based variant learns a molecular signature by weighting each pathway, we assessed the stability of this signature across the six breast cancer benchmarks. Each box in Figure IV.2B represents one of the 50 pathways used for the prediction task and are created from the six corresponding weights learned by the models trained on the different datasets. The pathway signature remains quite stable over the six different datasets and high (absolute) weights are associated with known cancer pathways like androgen response [264], hedgehog signaling [265], notch signaling [266], and MYC target [267]. With the introduction of attention-based COmic models, we introduce models with the capability of providing local interpretations, i.e., visualizations that provide insights into the decision process for a specific sample. We show an exemplary visualization of attention weights for three different, randomly chosen patients in the GSE11121 dataset in Figure IV.2C. For patient 1, the DMFS was correctly predicted to exceed 5 years. Patient 2 was correctly classified to have a DMFS below 5 years and patient 3 was wrongly classified to have a DMFS above 5 years while the actual DMFS of patient 3 was shorter than 5 years. The highest attention weights are associated with known cancer pathways. For patient 2, the highest amount of attention is given to hedgehog signaling. Androgen response gets the highest attention for patient 1 and 3. More examples can be found in the supplement.

One key advantage of artificial neural networks over kernel methods is their applicability on datasets with a vast number of samples. In the following, we will investigate, if our kernel networks provide the same applicability to large-scale datasets. Thus, we created simulated omics datasets of four different sizes: 100 samples, 1000 samples, 10000 samples, and 100000 samples. We then repeatedly trained pooling-based and attention-based COmic models on each simulated dataset five times and calculated the mean training time. Figure IV.2D shows the results. Since the batch size is usually chosen based on the number of samples in the training set, we calculated the mean training time for two different batch sizes. The blue and green lines show the training times of models trained with a fixed batch size of 32 samples per batch. The red and yellow lines show the training times of models with an adaptive batch size of one percent of the total sample count, i.e., each batch included a single sample, in case of the smallest simulated dataset, and 1000 samples , in case of the largest simulated dataset. The results show that COmic models can be easily trained on datasets with several hundreds of thousands of samples with the training time being linear depended on the number of samples. Furthermore, choosing an appropriate batch size can improve the training time by more than 50% on large-scale datasets.

IV.3.2 Multi-Omics Prediction on the METABRIC Benchmark Cohort

Since our proposed kernel layer can be incorporated into any ANN, COmic models can be flexibly expanded to multi-omics datasets. One possibility is to directly add the pathway kernels for the additional omics datatypes to the kernel layer, thereby increasing the number of graph Laplacians in the kernel layer. Another simple approach is to create sub-networks for each omics type, i.e., combine the output of pooling- or attention-based single-omics COmic models with a simple fully connected network. There are numerous other ways to expand COmic models to multi-omics data and, since our proposed approach is knowledge-driven, the individual solution has to be selected with the context of the data in mind. Similar to the authors of PIMKL, we chose the METABRIC cohort to investigate the practicality of applying COmic models to

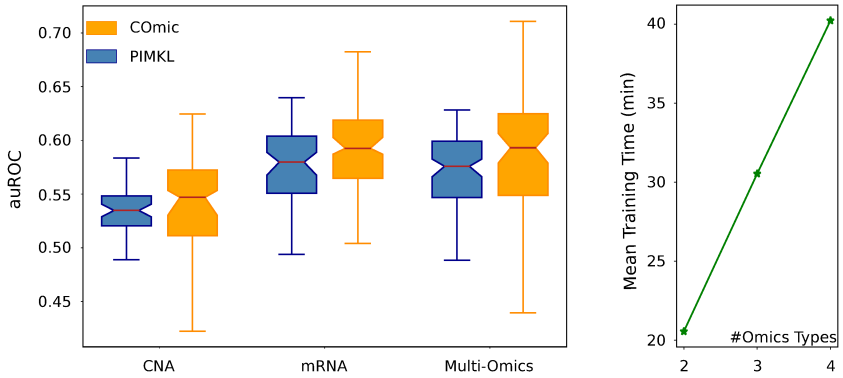

Figure IV.3: **Left**: Cross-validation performance of COmic and PIMKL models on the METABRIC cohort. The boxplots show the ten mean auROC validation scores of a 10-times repeated 10-fold cross-validation. Performances of COmic models are shown in orange. The center line of each box indicates the median. The height of the boxes represents the inter-quartile range (IQR) with the upper and lower whiskers set to 1.5 times the IQR. Notches represent the confidence interval (CI) around the median and were calculated using bootstrapping with 10000 iterations. **Right**: Mean training time of multi-omics COmic models. Artificial datasets with 2, 3, and 4 different omics modalities and 1000 data points per modality were investigated. Batch size and number of epochs were fixed as described for the METABRIC cohort. Training of models was repeated five times and mean times are indicated by stars.

multi-omics datasets. The METABRIC breast cancer cohort contains gene expression (mRNA) and copy number alteration (CNA) data. We performed the same prediction task as in [55], i.e., using molecular measures to predict whether a patient had recurrent cancer.

While we used pooling-based models with the same hyperparameters as described above for the single-omics prediction experiments, the creation of a multi-omics COmic model for the METABRIC cohort had to be done with caution. The issue with CNA data is that this datatype is tremendously sparse. To compensate for this sparseness, we build a network that used pooling-based kernel layers to compute an embedding for each datatype that is robust enough for sparse data and afterwards used a gated attention layer as described in section IV.2.3 to make the prediction. The pooling-based kernel layers used the same hyperparameters as the pooling-based COmic models in the single-omics experiments. For the gated attention layer we used an attention dimensionality of 4. Networks were trained for 200 epochs with the Adam optimizer [261] using the class-balanced loss function [262]. The batch size was set to 32. We used the auROC as our performance measure to enable the comparison to previously published PIMKL results on the METABRIC cohort. Additionally, we investigated the computational efficiency of COmic with regard to an increasing number of omics modalities.

The results of our experiments on the METABRIC cohort can be found on the left side of Figure IV.3. The shown PIMKL performance is taken from [55]. As expected, neither PIMKL nor COmic achieved good performance on the sparse CNA data. COmic models slightly outperformed PIMKL on the single-omics prediction task using gene expression data. While PIMKL shows a slightly decreased performance on the multi-omics prediction task, COmic seems to have the same performance on the multi-omics data as on the gene expression data

alone with an increase in variance. The runtime analysis (right side of Figure IV.3) shows a linear dependency on the number of omics modalities.

IV.4 Discussion

Kernel methods allow to induce prior knowledge into a prediction task resulting in increased robustness and the introduction of interpretation capabilities. In this work we propose COmic, a method to incorporate pathway-induced kernel functions into convolutional kernel networks. We are able to create learning models that can be robustly trained on small-scale datasets and scale very well with the number of samples. Thus, they can be efficiently applied to large datasets with hundreds of thousands of samples. Furthermore, our models provide global and local interpretations of predictions made on molecular measures due to the pathway-induced kernel function.

We used six different breast cancer cohorts to compare the performance of COmic models to previously proposed methods that use prior knowledge for prediction tasks with molecular measures as input data. The results presented in Figure IV.2A show that our method reaches state-of-the-art performance on classifying patients based on their DMFS/RFS from gene expression data: Compared to the considered competitors, COmic performs similar or even better. However, COmic models have the advantage that the time needed to train a model scales linearly with the number of samples (see Figure IV.2D). This enables the use of COmic models on datasets with hundreds of thousands of samples. We provide evidence that our method can be readily applied on large datasets by training models on simulated single-omics data with sizes ranging from 100 to 100,000 samples. Although datasets and patient cohorts used in computational biology and medicine traditionally have smaller sample sizes, high-throughput methods and the nowadays more frequently used big data paradigm will result in increasing sample counts in biological and medical datasets. At this day, TCGA already contains data from more than 85,000 patients. While methods that can deal with large datasets are usually deployed as black-box models, our method provides increased insight into the decision making process.

Using single-omics datasets strongly limits the decision process for diagnosis of a majority of diseases. Nowadays, it is well known that multi-omics information has to be incorporated to get a complete image of the pathomechanism causing a certain disease. Therefore, methods that are limited to a single datatype face serious constraints if employed as a decision-support system or to deepen knowledge about a pathomechanism. Our proposed method does not face the limitation of only using single-omics data as we show in our experiment with the METABRIC multi-omics cohort. The results show again that our method improves single-omics prediction as demonstrated by the performance on the gene expression data. The lower performance that both methods, PIMKL and COmic, show on the copy number alteration data can be explained by the sparseness of CNA data. Sparse data poses serious problems for prediction models [268] and both methods are not specifically designed for sparse data. However, we can show in Figure IV.3 that COmic models are able to achieve slightly better performance than PIMKL models on the multi-omics prediction task. This indicates that our approach could be advantageously used on multi-omics data, while the flexibility of the architecture (as described in section IV.3.2) enables researchers to tailor COmic models for specific datasets using domain

expertise.

Computing an interpretation of a machine learning method, either with *post-hoc* methods or through intrinsically interpretable models, is beneficial if and only if the interpretation serves a purpose. This purpose cannot be defined in general as it is highly dependent on the task, the data, and the user that is presented with the obtained interpretation. For the presented experiments, we investigated if the inherent interpretation capabilities of our COmic models are able to learn biological meaningful concepts directly from data. First, we considered the global interpretation capabilities of COmic. Here, COmic models assign a weight to each of the used pathways and the weights reflect the role that each pathway plays in classifying an input sample, i.e., a patient. We trained COmic models on six different single-omics breast cancer cohorts. Since we expect the biological processes in the cohorts to share high similarities, the weight signatures of all models should be similar if the COmic method is able to learn meaningful pathways from data. As shown in Figure IV.2B, this assumption is indeed well fulfilled with all six models having similar weight signatures. Furthermore, pathways with a high weight assigned to them are mainly known cancer-related pathways. Therefore, COmic models are able to learn biological meaningful pathway weights. Although the previously published PIMKL method also has a global interpretation capability, our method is able to learn pathways that are important for both, the negative and the positive class, due to the fact that the learned weights can be positive or negative. PIMKL only learns positive weights.

While global interpretation is useful to gain insights into a dataset, local interpretation can be used to get insights into the decision a model makes for a specific input. Attention-based COmic models can provide this insight utilizing the attention weights that are computed for each input sample separately. These weights directly determine the influence that each pathway has on the decision made by the model. We can visualize these influence using a heatmap (as shown in Figure IV.2C) to quickly see which pathways played an important role in the decision made. We randomly selected three samples from the GSE11121 dataset to evaluate if the attention weights are biological meaningful. Similar to the weight signatures of the globally interpretable COmic models, the attention weights of the locally interpretable COmic models highlighted known cancer-related pathways. Interestingly, the selected patient with a DMFS below five years has attention weights that are strongly focused on a single pathway. This is true for all correctly classified patients with a DFMS below five years (see supplement). On the other hand, patients with a DMFS below five years that were wrongly classified to have a DMFS above five years show attention weight patterns similar to those of patients with a DMFS above five years (see patient 3 in Figure IV.2 and additional examples in the supplement). This could indicate that the wrongly classified patients exhibit a different mechanism causing a DMFS below five years, compared to the correctly classified ones, which was not learned by the model. The local interpretation capabilities of COmic models can help to directly show possible directions to further investigate the data. Furthermore, the results of our experiments strongly suggest that both COmic model types are able to generate biologically meaningful interpretations. We chose heatmaps to visualize attention weights, since it appeared convenient for the considered prediction task on the studied dataset. However, different forms of explanations can be computed with attention weights, e.g., counterfactual explanation [269] and adversarial explanation [270]. The most suitable form of explanation is highly dependent on the application, target user group, and the goal aimed at by the explanation. Therefore, the chosen visualization should be understood as an example and not a general application

recommendation.

COmic models have a few hyperparameters that can be optimized using appropriate methods like, e.g., grid search or random search. These hyperparameters include the number of anchor points, the attention type, the dimensionality of the attention layer's parameters V and U, and the choice of pathways used for kernel layers. Furthermore, different initialization procedures for the anchor points can be explored, e.g., a parameter-free clustering that combines initializing anchor points with optimizing the number of anchor points for each pathway-induced kernel layer. We recommend to explore hyperparameter optimization when applying COmic models. However, minimizing energy consumption is a pressing concern that should be considered in every line of research nowadays. Therefore, we limited the computations performed for this work to the minimum required to support our claims. The hyperparameters for all models presented in this work were chosen by combining prior experience about kernel networks with domain expertise. Interestingly, this computation-free approach to hyperparameter selection already leads to competitive performance of our method on the considered prediction tasks.

IV.5 Conclusion

The introduced convolutional omics kernel networks utilize prior knowledge by pathway-induced kernel functions to provide robust end-to-end learning on small- to large-scale molecular measure datasets. Furthermore, utilizing pathway-induced kernel functions makes our method intrinsically interpretable with the ability to provide global and local interpretations.

We show the competitive performance of our method on six different single-omics breast cancer cohorts while providing new interpretation capabilities that exceed the possibilities of previously proposed methods. Furthermore, we show that COmic models can be readily adapted to multi-omics datasets.

On a larger scale, we show that incorporating a carefully crafted kernel function into an artificial neural network allows to robustly train ANNs on small-scale datasets as they frequently occur in computational biology and medicine. On the other hand, our method enables scientist to utilize kernel functions for large datasets as they arise more frequently with the increasing use of high-throughput methods and big data.

Acknowledgement

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. This research was supported by the German Federal Ministry of Education and Research (BMBF) project 'Training Center Machine Learning, Tübingen' with grant number 01|S17054. This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A.

Bibliography

[1] Alan Baker. Simplicity. *Stanford Encyclopedia of Philosophy*, 2004.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873): 583–589, 2021.

[4] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *CoRR*, abs/2102.12092, 2021. URL https://arxiv.org/abs/2102.12092.

[5] VN Vapnik. Principles of risk minimization for learning theory, advances in neural information processing nips 4 (pp. 831±838), 1992.

[6] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[7] Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pages 10767–10777. PMLR, 2020.

[8] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[9] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

[10] Marilyn Strathern. 'improving ratings': audit in the british university system. *European review*, 5(3):305–321, 1997.

[11] Charles AE Goodhart and CAE Goodhart. *Problems of monetary management: the UK experience*. Springer, 1984.

[12] Herbert A Simon. Spurious correlation: A causal interpretation. *Journal of the American statistical Association*, 49(267):467–479, 1954.

[13] Brian D Haig. What is a spurious correlation? *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, 2(2):125–132, 2003.

[14] Robert S Witte and John S Witte. *Statistics*. John Wiley & Sons, 2017.

[15] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[16] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):1–28, 2015.

[17] Tom Fawcett. Roc graphs: Notes and practical considerations for researchers. *Machine learning*, 31(1):1–38, 2004.

[18] Toon Calders and Szymon Jaroszewicz. Efficient auc optimization for classification. In *Knowledge Discovery in Databases: PKDD 2007: 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17-21, 2007. Proceedings 11*, pages 42–53. Springer, 2007.

[19] Steve Halligan, Douglas G Altman, and Susan Mallett. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *European radiology*, 25:932–939, 2015.

[20] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[21] Mu Zhu. Recall, precision and average precision. *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo*, 2(30):6, 2004.

[22] James Briggs. Web-Scraping and Pre-Processing for NLP - Towards Data Science, Dec 2021. URL https://towardsdatascience.com/web-scraping-and-pre-processing-for-nlp-2e78810b40f1. Accessed: 2023, February 22.

[23] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. Documenting computer vision datasets: an invitation to reflexive data practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 161–172, 2021.

[24] Haotian Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, and Liwei Wang. Towards a theoretical framework of out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:23519–23531, 2021.

[25] George Box. Signal-to-noise ratios, performance criteria, and transformations. *Technometrics*, 30(1):1–17, 1988.

[26] Anders Krogh, Björn Larsson, Gunnar Von Heijne, and Erik LL Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of molecular biology*, 305(3):567–580, 2001.

[27] Jeppe Hallgren, Konstantinos D Tsirigos, Mads Damgaard Pedersen, José Juan Almagro Armenteros, Paolo Marcatili, Henrik Nielsen, Anders Krogh, and Ole Winther. Deeptmhmm predicts alpha and beta transmembrane proteins using deep neural networks. *bioRxiv*, 2022.

[28] José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

[29] Vineet Thumuluri, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Henrik Nielsen, and Ole Winther. Deeploc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Research*, 2022.

[30] Lukas Käll, Anders Krogh, and Erik LL Sonnhammer. Advantages of combined transmembrane topology and signal peptide prediction—the phobius web server. *Nucleic acids research*, 35(suppl_2):W429–W432, 2007.

[31] C Safran. Update on data reuse in health care. *Yearbook of medical informatics*, 26(01): 24–27, 2017.

[32] World Health Organization et al. Sharing and reuse of health-related data for research purposes: Who policy and implementation guidance. 2022.

[33] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[34] Chief Ben-Eghan, Rosie Sun, Jose Sergio Hleap, Alex Diaz-Papkovich, Hans Markus Munter, Audrey V Grant, Charles Dupras, and Simon Gravel. Don't ignore genetic data from minority populations. *Nature*, 585(7824):184–186, 2020.

[35] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica*, 23:77–91, 2016.

[36] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23): 12592–12594, 2020.

[37] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[38] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[39] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[40] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[41] Richard Bellman. Dynamic programming. *science*, 153(3731):34–37, 1966.

[42] Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Measures of complexity: festschrift for alexey chervonenkis*, pages 11–30, 2015.

[43] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *Computational Learning Theory: 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001 Amsterdam, The Netherlands, July 16–19, 2001 Proceedings 14*, pages 416–426. Springer, 2001.

[44] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[45] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[46] Christina Leslie, Eleazar Eskin, and William Stafford Noble. The spectrum kernel: a string kernel for svm protein classification. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 564—575, 2002. ISSN 2335-6928. URL http://europepmc.org/abstract/MED/11928508.

[47] Christina Leslie and Rui Kuang. Fast string kernels using inexact matching for protein sequences. *Journal of Machine Learning Research*, 5(Nov):1435–1455, 2004.

[48] Eleazar Eskin, Jason Weston, William S Noble, and Christina S Leslie. Mismatch string kernels for svm protein classification. In *Advances in neural information processing systems*, pages 1441–1448, 2003.

[49] G. Rätsch, S. Sonnenburg, and B. Schölkopf. Rase: recognition of alternatively spliced exons in c.elegans. *Bioinformatics*, 21(suppl_1):369–377, 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti1053. URL https://doi.org/10.1093/bioinformatics/bti1053.

[50] Hisashi Kashima, Koji Tsuda, and Akihiro Inokuchi. Kernels for graphs. In *Kernel methods in computational biology*, pages 155–170. MIT Press, 2004.

[51] Pierre Mahé, Nobuhisa Ueda, Tatsuya Akutsu, Jean-Luc Perret, and Jean-Philippe Vert. Graph kernels for molecular structure- activity relationship analysis with support vector machines. *Journal of chemical information and modeling*, 45(4):939–951, 2005.

[52] S Joshua Swamidass, Jonathan Chen, Jocelyne Bruand, Peter Phung, Liva Ralaivola, and Pierre Baldi. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics*, 21(suppl_1):i359–i368, 2005.

[53] Peter Meinicke, Maike Tech, Burkhard Morgenstern, and Rainer Merkl. Oligo kernels for datamining on biological sequences: a case study on prokaryotic translation initiation sites. *BMC bioinformatics*, 5(1):169, 2004.

[54] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4): 395–416, 2007.

[55] Matteo Manica, Joris Cadow, Roland Mathis, and Maria Rodriguez Martinez. Pimkl: pathway-induced multiple kernel learning. *NPJ systems biology and applications*, 5(1): 1–8, 2019.

[56] Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, 6 (1):1–25, 2019.

[57] Martin R Cowie, Juuso I Blomster, Lesley H Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, et al. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106:1–9, 2017.

[58] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[59] Tobias Glasmachers. Limits of end-to-end learning. In *Asian conference on machine learning*, pages 17–32. PMLR, 2017.

[60] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

[61] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

[62] Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. The rise of deep learning in drug discovery. *Drug discovery today*, 23(6):1241–1250, 2018.

[63] Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, 11(11):5088, 2021.

[64] Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. *Advances in neural information processing systems*, 22:342–350, 2009.

[65] Julien Mairal, Piotr Koniusz, Zaid Harchaoui, and Cordelia Schmid. Convolutional kernel networks. In *Advances in neural information processing systems*, pages 2627–2635, 2014.

[66] Julien Mairal. End-to-end kernel learning with supervised convolutional kernel networks. In *Advances in neural information processing systems*, pages 1399–1407, 2016.

[67] Dexiong Chen, Laurent Jacob, and Julien Mairal. Biological sequence modeling with convolutional kernel networks. *Bioinformatics*, 35(18):3294–3302, 2019.

[68] Dexiong Chen, Laurent Jacob, and Julien Mairal. Recurrent kernel networks. In *Advances in Neural Information Processing Systems*, pages 13431–13442, 2019.

[69] Dexiong Chen, Laurent Jacob, and Julien Mairal. Convolutional kernel networks for graph-structured data. In *International Conference on Machine Learning*, pages 1576–1586. PMLR, 2020.

[70] Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Proceedings of the 14th annual conference on neural information processing systems*, number CONF, pages 682–688, 2001.

[71] Kai Zhang, Ivor W Tsang, and James T Kwok. Improved nyström low-rank approximation and error analysis. In *Proceedings of the 25th international conference on Machine learning*, pages 1232–1239, 2008.

[72] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[73] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[74] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[75] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[76] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.

[77] Ilia Stepin, Jose M Alonso, Alejandro Catala, and Martín Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021.

[78] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019.

[79] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222, 2017.

[80] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[81] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

[82] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[83] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215, 2019.

[84] Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified bp attributions fail. In *International Conference on Machine Learning*, pages 9046–9057. PMLR, 2020.

[85] Sebastian Bordt, Michèle Finck, Eric Raidl, and Ulrike von Luxburg. Post-hoc explanations fail to achieve their purpose in adversarial contexts. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 891–905, 2022.

[86] Jan F. Veneman, Rik Kruidhof, Edsko E. G. Hekman, Ralf Ekkelenkamp, Edwin H. F. Van Asseldonk, and Herman van der Kooij. Design and evaluation of the lopes exoskeleton robot for interactive gait rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 15(3):379–386, 2007. doi: 10.1109/TNSRE.2007.903919.

[87] Sue VG Cobb, Sarah Nichols, Amanda Ramsey, and John R Wilson. Virtual reality-induced symptoms and effects (vrise). *Presence: Teleoperators & Virtual Environments*, 8(2):169–186, 1999.

[88] Shayan Jalilpour and Gernot Müller-Putz. Toward passive bci: asynchronous decoding of neural responses to direction-and angle-specific perturbations during a simulated cockpit scenario. *Scientific Reports*, 12(1):6802, 2022.

[89] World Health Organization et al. *World malaria report 2021*. World Health Organization, 2021.

[90] World Health Organization et al. *World malaria report 2022*. World Health Organization, 2022.

[91] Prasanna Jagannathan and Abel Kakuru. Malaria in 2022: Increasing challenges, cautious optimism. *Nature communications*, 13(1):1–3, 2022.

[92] Benjamin Mordmüller, Güzin Surat, Heimo Lagler, Sumana Chakravarty, Andrew S Ishizuka, Albert Lalremruata, Markus Gmeiner, Joseph J Campo, Meral Esen, Adam J Ruben, et al. Sterile protection against human malaria by chemoattenuated pfspz vaccine. *Nature*, 542(7642):445–449, 2017.

[93] Matthias Döring, Joachim Büch, Georg Friedrich, Alejandro Pironti, Prabhav Kalaghatgi, Elena Knops, Eva Heger, Martin Obermeier, Martin Däumer, Alexander Thielen, et al. geno2pheno [ngs-freq]: a genotypic interpretation system for identifying viral drug resistance using next-generation sequencing data. *Nucleic acids research*, 46(W1):W271–W277, 2018.

[94] Wenyi Yang and Lei Deng. Predba: A heterogeneous ensemble approach for predicting protein-dna binding affinity. *Scientific Reports*, 10(1):1–11, 2020.

[95] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome biology*, 18(1):1–15, 2017.

[96] Mike X Cohen. *Analyzing neural time series data: theory and practice*. MIT press, 2014.

[97] Christian Kothe. Lab streaming layer (lsl), 2020. https://github.com/sccn/labstreaminglayer.

[98] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[99] Kristian E Swearingen, Scott E Lindner, Lirong Shi, Melanie J Shears, Anke Harupa, Christine S Hopp, Ashley M Vaughan, Timothy A Springer, Robert L Moritz, Stefan HI Kappe, et al. Interrogating the plasmodium sporozoite surface: identification of surface-exposed proteins and demonstration of glycosylation on csp and trap by mass spectrometry-based proteomics. *PLoS pathogens*, 12(4):e1005606, 2016.

[100] Franz Baumdicker, Wolfgang R Hess, and Peter Pfaffelhuber. The infinitely many genes model for the distributed genome of bacteria. *Genome biology and evolution*, 4(4):443–456, 2012.

[101] Wei Ding, Franz Baumdicker, and Richard A Neher. panx: pan-genome analysis and exploration. *Nucleic acids research*, 46(1):e5–e5, 2018.

[102] Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst. Learning laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 64(23):6160–6173, 2016.

[103] Vassilis Kalofolias. How to learn a graph from smooth signals. In *Artificial Intelligence and Statistics*, pages 920–929. PMLR, 2016.

[104] Hichem Sahbi. Learning laplacians in chebyshev graph convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2064–2075, 2021.

[105] Tao Lei, Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Deriving neural architectures from sequence and graph kernels. *arXiv preprint arXiv:1705.09037*, 2017.

[106] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL https://aclanthology.org/P19-1355.

[107] Yann LeCun, Leon Bottou, Genevieve B Orr, Klaus-Robert Müller, et al. Neural networks: Tricks of the trade. *Springer Lecture Notes in Computer Sciences*, 1524(5-50):6, 1998.

[108] Steven M LaValle, Michael S Branicky, and Stephen R Lindemann. On the relationship between classical grid search and probabilistic roadmaps. *The International Journal of Robotics Research*, 23(7-8):673–692, 2004.

[109] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[110] Jason Hickel. *Less is more: How degrowth will save the world*. Random House, 2020.

[111] Jessica Hernandez. *Fresh banana leaves: Healing Indigenous landscapes through Indigenous science*. North Atlantic Books, 2022.

[112] Ashish Kothari, Federico Demaria, and Alberto Acosta. Buen vivir, degrowth and ecological swaraj: Alternatives to sustainable development and the green economy. *Development*, 57(3-4):362–375, 2014.

[113] Matthias Schwab and Elke Schaeffeler. Pharmacogenomics: a key component of personalized therapy. *Genome Medicine*, 4(11):1–3, 2012.

[114] Ulrich M Zanger and Matthias Schwab. Cytochrome p450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & therapeutics*, 138(1):103–141, 2013.

[115] Kathrin Klein, Roman Tremmel, Stefan Winter, Sarah Fehr, Florian Battke, Tim Scheurenbrand, Elke Schaeffeler, Saskia Biskup, Matthias Schwab, and Ulrich M Zanger. A new panel-based next-generation sequencing method for adme genes reveals novel associations of common and rare variants with expression in a human liver cohort. *Frontiers in Genetics*, 10:7, 2019.

[116] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[117] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, page 2801–2807. AAAI Press, 2019. ISBN 9780999241141.

[118] Future of Life Institute. "pause giant ai experiments: An open letter", March 22, 2023.

[119] European Commission. *Laying Down Harmonizing Rules of Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. 2021.

[120] DIN and DKE. *Deutsche Normungsroadmap Künstliche Intelligenz (Ausgabe 2)*. 2022.

[121] Judea Pearl. *Causality*. Cambridge university press, 2009.

[122] Stephen L Morgan and Christopher Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2015.

[123] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

[124] Gianluca Bontempi and Maxime Flauder. From dependency to causality: a machine learning approach. *Journal of Machine Learning Research*, 16(1):2437–2457, 2015.

[125] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[126] Bernhard Schölkopf. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 765–804. 2022.

[127] Sara Mohammad-Taheri, Jeremy Zucker, Charles Tapley Hoyt, Karen Sachs, Vartika Tewari, Robert Ness, and Olga Vitek. Do-calculus enables estimation of causal effects in partially observed biomolecular pathways. *Bioinformatics*, 38(Supplement_1):i350–i358, 2022.

[128] V Dietz, J Quintern, and W Berger. Cerebral evoked potentials associated with the compensatory reactions following stance and gait perturbation. *Neuroscience letters*, 50 (1-3):181–186, 1984.

[129] H Ackermann, HC Diener, and J Dichgans. Mechanically evoked cerebral potentials and long-latency muscle responses in the evaluation of afferent and efferent long-loop pathways in humans. *Neuroscience letters*, 66(3):233–238, 1986.

[130] RB Duckrow, K Abu-Hasaballah, R Whipple, and L Wolfson. Stance perturbation-evoked potentials in old people with poor gait and balance. *Clinical Neurophysiology*, 110(12): 2026–2032, 1999.

[131] Richard W. Staines, William E. McIlroy, and John D. Brooke. Cortical representation of whole-body movement is modulated by proprioceptive discharge in humans. *Experimental Brain Research*, 138:235–42, 2001.

[132] Allan L Adkin, Sylvia Quant, Brian E Maki, and William E McIlroy. Cortical responses associated with predictable and unpredictable compensatory balance reactions. *Experimental Brain Research*, 172:85–93, 2006.

[133] Jessy Parokaran Varghese, Robert E. McIlroy, and Michael Barnett-Cowan. Perturbation-evoked potentials: Significance and application in balance control research. *Neuroscience & Biobehavioral Reviews*, 83:267–280, 2017. ISSN 0149-7634. doi: https://doi.org/10.1016/j.neubiorev.2017.10.022. URL https://www.sciencedirect.com/science/article/pii/S0149763417305699.

[134] B Dimitrov, T Gavrilenko, and P Gatev. Mechanically evoked cerebral potentials to sudden ankle dorsiflexion in human subjects during standing. *Neuroscience letters*, 208 (3):199–202, 1996.

[135] Amanda Marlin, George Mochizuki, William R Staines, and William E McIlroy. Localizing evoked cortical activity associated with balance reactions: does the anterior cingulate play a role? *Journal of neurophysiology*, 111(12):2634–2643, 2014.

[136] Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain–computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002. ISSN 1388-2457. doi: https://doi.org/10.1016/S1388-2457(02)00057-3. URL https://www.sciencedirect.com/science/article/pii/S1388245702000573.

[137] José del R Millán, Rüdiger Rupp, Gernot Mueller-Putz, Roderick Murray-Smith, Claudio Giugliemma, Michael Tangermann, Carmen Vidaurre, Febo Cincotti, Andrea Kubler, Robert Leeb, et al. Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges. *Frontiers in neuroscience*, page 161, 2010.

[138] Jonathan Wolpaw and Elizabeth Winter Wolpaw. *Brain–Computer Interfaces: Principles and Practice*. Oxford University Press, 01 2012. ISBN 9780195388855. doi: 10.1093/acprof:oso/9780195388855.001.0001. URL https://doi.org/10.1093/acprof:oso/9780195388855.001.0001.

[139] Nikunj A Bhagat, Anusha Venkatakrishnan, Berdakh Abibullaev, Edward J Artz, Nuray Yozbatiran, Amy A Blank, James French, Christof Karmonik, Robert G Grossman, Marcia K O'Malley, et al. Design and optimization of an eeg-based brain machine interface (bmi) to an upper-limb exoskeleton for stroke survivors. *Frontiers in neuroscience*, 10:122, 2016.

[140] Simona Crea, Marius Nann, Emilio Trigili, Francesca Cordella, Andrea Baldoni, Francisco Javier Badesa, José Maria Catalán, Loredana Zollo, Nicola Vitiello, Nicolas Garcia Aracil, et al. Feasibility and safety of shared eeg/eog and vision-guided autonomous whole-arm exoskeleton control to perform activities of daily living. *Scientific reports*, 8 (1):10823, 2018.

[141] T. O. Zander, C. Kothe, S. Welke, and M. Roetting. Utilizing secondary input from passive brain–computer interfaces for enhancing human–machine interaction. In *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience (Berlin: Springer)*, pages 759–71, 2009. doi: 10.1007/978-3-642-02812-0_86. URL https://doi.org/10.1007/978-3-642-02812-0_86.

[142] Thorsten O Zander and Christian Kothe. Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. *Journal of Neural Engineering*, 8(2):025005, 2011. doi: 10.1088/1741-2560/8/2/025005. URL https://dx.doi.org/10.1088/1741-2560/8/2/025005.

[143] Peter Gerjets, Carina Walter, Wolfgang Rosenstiel, Martin Bogdan, and Thorsten O Zander. Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Frontiers in neuroscience*, 8:385, 2014.

[144] Marten K Scheffers and Michael GH Coles. Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors. *Journal of Experimental Psychology: Human Perception and Performance*, 26(1):141, 2000.

[145] Lucas C Parra, Clay D Spence, Adam D Gerson, and Paul Sajda. Response error correction-a demonstration of improved human-machine performance using real-time eeg

monitoring. *IEEE transactions on neural systems and rehabilitation engineering*, 11(2): 173–177, 2003.

[146] Catarina Lopes Dias, Andreea I Sburlea, and Gernot R Müller-Putz. Masked and unmasked error-related potentials during continuous control and feedback. *Journal of neural engineering*, 15(3):036031, 2018.

[147] Hiroshi Shibasaki and Mark Hallett. What is the bereitschaftspotential? *Clinical Neurophysiology*, 117(11):2341–2356, 2006. ISSN 1388-2457. doi: https://doi.org/10.1016/j.clinph.2006.04.025. URL https://www.sciencedirect.com/science/article/pii/S138824570600229X.

[148] Matthias Schultze-Kraft, Daniel Birman, Marco Rusconi, Carsten Allefeld, Kai Görgen, Sven Dähne, Benjamin Blankertz, and John-Dylan Haynes. The point of no return in vetoing self-initiated movements. *Proceedings of the national Academy of Sciences*, 113 (4):1080–1085, 2016.

[149] Josef Faller, Carmen Vidaurre, Teodoro Solis-Escalante, Christa Neuper, and Reinhold Scherer. Autocalibration and recurrent adaptation: Towards a plug and play online erd-bci. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(3): 313–319, 2012.

[150] Andreas Schwarz, Reinhold Scherer, David Steyrl, Josef Faller, and Gernot R Müller-Putz. A co-adaptive sensory motor rhythms brain-computer interface based on common spatial patterns and random forest. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1049–1052. IEEE, 2015.

[151] Miguel Angel Ortiz Pérez and R Benjamin Knapp. *BioTools: a biosignal toolbox for composers and performers*. 2008.

[152] Andreas Schwarz, Patrick Ofner, Joana Pereira, Andreea Ioana Sburlea, and Gernot R Müller-Putz. Decoding natural reach-and-grasp actions from human eeg. *Journal of Neural Engineering*, 15(1):016005, dec 2017.

[153] Andreas Schwarz, Joana Pereira, Reinmar Kobler, and Gernot R Müller-Putz. Unimanual and bimanual reach-and-grasp actions can be decoded from human eeg. *IEEE transactions on biomedical engineering*, 67(6):1684–1695, 2019.

[154] Jerome H Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

[155] Benjamin Blankertz, Steven Lemm, Matthias Treder, Stefan Haufe, and Klaus-Robert Müller. Single-trial analysis and classification of erp components—a tutorial. *NeuroImage*, 56(2):814–825, 2011.

[156] Gernot Müller-Putz, Reinhold Scherer, Clemens Brunner, Robert Leeb, and Gert Pfurtscheller. Better than random: a closer look on bci results. *International journal of bioelectromagnetism*, 10:52–55, 2008.

[157] Martin Billinger, Ian Daly, Vera Kaiser, Jing Jin, Brendan Z Allison, Gernot R Müller-Putz, and Clemens Brunner. *Is it significant? Guidelines for reporting BCI performance*. Springer, 2013.

[158] G Mochizuki, KM Sibley, JG Esposito, JM Camilleri, and WE McIlroy. Cortical responses associated with the preparation and reaction to full-body perturbations to upright stability. *Clinical Neurophysiology*, 119(7):1626–1637, 2008.

[159] V Dietz, J Quintern, W Berger, and E Schenck. Cerebral potentials and leg muscle emg responses associated with stance perturbation. *Experimental Brain Research*, 57(2): 348–354, 1985.

[160] V Dietz, J Quintern, and W Berger. Afferent control of human stance and gait: evidence for blocking of group i afferents during gait. *Experimental Brain Research*, 61:153–163, 1985.

[161] Teodoro Solis-Escalante, Joris van der Cruijsen, Digna de Kam, Joost van Kordelaar, Vivian Weerdesteyn, and Alfred C. Schouten. Cortical dynamics during preparation and execution of reactive balance responses with distinct postural demands. *NeuroImage*, 188: 557–571, 2019. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2018.12.045. URL https://www.sciencedirect.com/science/article/pii/S105381191832189X.

[162] George Mochizuki, Kathryn M Sibley, Hannah J Cheung, Joanne M Camilleri, and William E McIlroy. Generalizability of perturbation-evoked cortical potentials: independence from sensory, motor and overall postural state. *Neuroscience letters*, 451(1):40–44, 2009.

[163] S Quant, Allan L Adkin, WR Staines, and WE McIlroy. Cortical activation following a balance disturbance. *Experimental brain research*, 155:393–400, 2004.

[164] Franklin F Offner. The eeg as potential mapping: the value of the average monopolar reference. *Electroencephalography and clinical neurophysiology*, 2(2):213–214, 1950.

[165] JW Osselton. Acquisition of eeg data by bipolar unipolar and average reference methods: a theoretical comparison. *Electroencephalography and clinical neurophysiology*, 19(5): 527–528, 1965.

[166] Sylvia Quant, Brian E Maki, and William E McIlroy. The association between later cortical potentials and later phases of postural reactions evoked by perturbations to upright stance. *Neuroscience letters*, 381(3):269–274, 2005.

[167] Kathryn M. Sibley, George Mochizuki, James S. Frank, and William E. McIlroy. The relationship between physiological arousal and cortical and autonomic responses to postural instability. *Experimental Brain Research*, 203:533–40, 2010.

[168] Guangyu Bin, Xiaorong Gao, Zheng Yan, Bo Hong, and Shangkai Gao. An online multi-channel ssvep-based brain–computer interface using a canonical correlation analysis method. *Journal of Neural Engineering*, 6(4):046002, 2009.

[169] Janir Nuno da Cruz, Feng Wan, Chi Man Wong, and Teng Cao. Adaptive time-window length based on online performance measurement in ssvep-based bcis. *Neurocomputing*, 149:93–99, 2015.

[170] Catarina Lopes-Dias, Andreea I Sburlea, and Gernot R Müller-Putz. Online asynchronous decoding of error-related potentials during the continuous control of a robot. *Scientific reports*, 9(1):17596, 2019.

[171] Mads Jochumsen, Imran Khan Niazi, Kim Dremstrup, and Ernest Nlandu Kamavuako. Detecting and classifying three different hand movement types through electroencephalography recordings for neurorehabilitation. *Medical & biological engineering & computing*, 54:1491–1501, 2016.

[172] Andreas Pinegger, Josef Faller, Sebastian Halder, Selina C Wriessnegger, and Gernot R Müller-Putz. Control or non-control state: that is the question! an asynchronous visual p300-based bci approach. *Journal of neural engineering*, 12(1):014001, 2015.

[173] Eleanor M Riley and V Ann Stewart. Immune mechanisms in malaria: new insights in vaccine development. *Nature medicine*, 19(2):168, 2013.

[174] Henry M Wu. Evaluation of the sick returned traveler. In *Seminars in diagnostic pathology*, volume 36, pages 197–202. Elsevier, 2019.

[175] SCTP Rts. Efficacy and safety of rts, s/as01 malaria vaccine with or without a booster dose in infants and children in africa: final results of a phase 3, individually randomised, controlled trial. *The Lancet*, 386(9988):31–45, 2015.

[176] Ally Olotu, Gregory Fegan, Juliana Wambua, George Nyangweso, Ken O Awuondo, Amanda Leach, Marc Lievens, Didier Leboulleux, Patricia Njuguna, Norbert Peshu, et al. Four-year efficacy of rts, s/as01e and its interaction with malaria exposure. *New England Journal of Medicine*, 368(12):1111–1120, 2013.

[177] Joshua M Obiero, Joseph J Campo, Anja Scholzen, Arlo Randall, Else M Bijker, Meta Roestenberg, Cornelus C Hermsen, Andy Teng, Aarti Jain, D Huw Davies, et al. Antibody biomarkers associated with sterile protection induced by controlled human malaria infection under chloroquine prophylaxis. *Msphere*, 4(1):e00027–19, 2019.

[178] Monya Baker. Making membrane proteins for structures: a trillion tiny tweaks. *Nature methods*, 7(6):429–434, 2010.

[179] Renu Tuteja. Malaria- an overview. *The FEBS journal*, 274(18):4670–4679, 2007.

[180] Mats Wahlgren, Suchi Goel, and Reetesh R Akhouri. Variant surface antigens of plasmodium falciparum and their roles in severe malaria. *Nature Reviews Microbiology*, 15(8):479–491, 2017.

[181] Beatrice Amos, Cristina Aurrecoechea, Matthieu Barba, Ana Barreto, Evelina Y Basenko, Robert Belnap, Ann S Blevins, Ulrike Böhme, John Brestelli, Brian P Brunk, et al. Veupathdb: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Research*, 50(D1):D898–D911, 2022.

[182] Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters. The immune epitope database (iedb): 2018 update. *Nucleic acids research*, 47(D1): D339–D343, 2019.

[183] Andrew R Osborne, Kaye D Speicher, Pamela A Tamez, Souvik Bhattacharjee, David W Speicher, and Kasturi Haldar. The host targeting motif in exported plasmodium proteins is cleaved in the parasite endoplasmic reticulum. *Molecular and biochemical parasitology*, 171(1):25–31, 2010.

[184] Thorey K Jonsdottir, Mikha Gabriela, Brendan S Crabb, Tania F de Koning-Ward, and Paul R Gilson. Defining the essential exportome of the malaria parasite. *Trends in Parasitology*, 37(7):664–675, 2021.

[185] Malcolm J Gardner, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, Arnab Pain, Karen E Nelson, Sharen Bowman, et al. Genome sequence of the human malaria parasite plasmodium falciparum. *Nature*, 419(6906): 498–511, 2002.

[186] D AK, Deepti Shrivastava, Amogh A Sahasrabuddhe, Saman Habib, and Vishal Trivedi. Plasmodium falciparum fikk9. 1 is a monomeric serine-threonine protein kinase with features to exploit as a drug target. *Chemical Biology & Drug Design*, 2021.

[187] Sarah J Tarr, Robert W Moon, Iris Hardege, and Andrew R Osborne. A conserved domain targets exported phistb family proteins to the periphery of plasmodium infected erythrocytes. *Molecular and biochemical parasitology*, 196(1):29–40, 2014.

[188] Ankit Gupta, Girija Thiruvengadam, and Sanjay A Desai. The conserved clag multigene family of malaria parasites: essential roles in host–pathogen interaction. *Drug Resistance Updates*, 18:47–54, 2015.

[189] The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023.

[190] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[191] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

[192] Anders Krogh, Björn Larsson, Gunnar Von Heijne, and Erik LL Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of molecular biology*, 305(3):567–580, 2001.

[193] Lukas Zimmermann, Andrew Stephens, Seung-Zin Nam, David Rau, Jonas Kübler, Marko Lozajic, Felix Gabler, Johannes Söding, Andrei N Lupas, and Vikram Alva. A completely reimplemented mpi bioinformatics toolkit with a new hhpred server at its core. *Journal of molecular biology*, 430(15):2237–2243, 2018.

[194] Felix Gabler, Seung-Zin Nam, Sebastian Till, Milot Mirdita, Martin Steinegger, Johannes Söding, Andrei N Lupas, and Vikram Alva. Protein sequence analysis using the mpi bioinformatics toolkit. *Current Protocols in Bioinformatics*, 72(1):e108, 2020.

[195] Amos Bairoch, Rolf Apweiler, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, et al. The universal protein resource (uniprot). *Nucleic acids research*, 33(suppl_1): D154–D159, 2005.

[196] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[197] Martin Steinegger, Milot Mirdita, and Johannes Söding. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature methods*, 16(7): 603–606, 2019.

[198] Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

[199] Sven Degroeve, Bernard De Baets, Yves Van de Peer, and Pierre Rouzé. Feature subset selection for splice site prediction. *Bioinformatics*, 18(suppl_2):S75–S83, 2002.

[200] Alexander Zien, Gunnar Rätsch, Sebastian Mika, Bernhard Schölkopf, Thomas Lengauer, and K-R Müller. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, 16(9):799–807, 2000.

[201] Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*, 10 (1):1–14, 2019.

[202] Timor Kadir and Michael Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[203] Liefeng Bo, Kevin Lai, Xiaofeng Ren, and Dieter Fox. Object recognition with hierarchical kernel descriptors. In *CVPR 2011*, pages 1729–1736. IEEE, 2011.

[204] Grégoire Mialon, Dexiong Chen, Alexandre d'Aspremont, and Julien Mairal. A trainable optimal transport embedding for feature aggregation and its relationship to attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=ZK6vTvb84s.

[205] Soo-Yon Rhee, Matthew J Gonzales, Rami Kantor, Bradley J Betts, Jaideep Ravela, and Robert W Shafer. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic acids research*, 31(1):298–303, 2003.

[206] Robert W Shafer. Rationale and uses of a public hiv drug-resistance database. *The Journal of infectious diseases*, 194(Supplement_1):S51–S58, 2006.

[207] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2): 442–451, 1975.

[208] Letícia M Raposo, Paulo Tadeu CR Rosa, and Flavio F Nobre. Random forest algorithm for prediction of hiv drug resistance. In *Pattern Recognition Techniques Applied to Biomedical Problems*, pages 109–127. Springer, 2020.

[209] Sören Sonnenburg, Gabriele Schweikert, Petra Philips, Jonas Behr, and Gunnar Rätsch. Accurate splice site prediction using support vector machines. In *BMC bioinformatics*, volume 8, pages 1–16. Springer, 2007.

[210] Abdul KMA Baten, Bill CH Chang, Saman K Halgamuge, and Jason Li. Splice site identification using probabilistic parameters and svm classification. In *BMC bioinformatics*, volume 7, pages 1–15. BioMed Central, 2006.

[211] Jasper Zuallaert, Fréderic Godin, Mijung Kim, Arne Soete, Yvan Saeys, and Wesley De Neve. Splicerover: interpretable convolutional neural networks for improved splice site prediction. *Bioinformatics*, 34(24):4180–4188, 2018.

[212] Martin G Reese, Frank H Eeckman, David Kulp, and David Haussler. Improved splice site detection in genie. *Journal of computational biology*, 4(3):311–323, 1997.

[213] Te-Ming Chen, Chung-Chin Lu, and Wen-Hsiung Li. Prediction of splice sites with dependency graphs and their expanded bayesian networks. *Bioinformatics*, 21(4):471–482, 2005.

[214] Davide Chicco. Ten quick tips for machine learning in computational biology. *BioData mining*, 10(1):1–17, 2017.

[215] Soo-Yon Rhee, Jonathan Taylor, W Jeffrey Fessel, David Kaufman, William Towner, Paolo Troia, Peter Ruane, James Hellinger, Vivian Shirvani, Andrew Zolopa, et al. Hiv-1 protease mutations and protease inhibitor cross-resistance. *Antimicrobial agents and chemotherapy*, 54(10):4253–4261, 2010.

[216] Richard Colonno, Ronald Rose, Colin McLaren, Alexandra Thiry, Neil Parkin, and Jacques Friborg. Identification of i50l as the signature atazanavir (atv)-resistance mutation in treatment-naive hiv-1-infected patients receiving atv-containing regimens. *Journal of Infectious Diseases*, 189(10):1802–1810, 2004.

[217] JAAP Goudsmit, ANTHONY De Ronde, David D Ho, and Alan S Perelson. Human immunodeficiency virus fitness in vivo: calculations based on a single zidovudine resistance mutation at codon 215 of reverse transcriptase. *Journal of virology*, 70(8):5662–5664, 1996.

[218] Richard Bethell, Joseph Scherer, Myriam Witvrouw, Agnes Paquet, Eoin Coakley, and David Hall. Phenotypic protease inhibitor resistance and cross-resistance in the clinic from 2006 to 2008 and mutational prevalences in hiv from patients with discordant tipranavir and darunavir susceptibility phenotypes. *AIDS research and human retroviruses*, 28(9): 1019–1024, 2012.

[219] Brendan A Larder, Sharon D Kemp, and P Richard Harrigan. Potential mechanism for sustained antiretroviral efficacy of azt-3tc combination therapy. *Science*, 269(5224): 696–699, 1995.

[220] Jonathan M Schapiro, Mark A Winters, Fran Stewart, Bradley Efron, Jane Norris, Michael J Kozal, and Thomas C Merigan. The effect of high-dose saquinavir on viral load and cd4+ t-cell counts in hiv-infected patients. *Annals of Internal Medicine*, 124 (12):1039–1050, 1996.

[221] Charles Craig, Esther Race, Jonathan Sheldon, Lynne Whittaker, Sue Gilbert, Alec Moffatt, Jane Rose, Shobana Dissanayeke, Gung-Wei Chirn, Ian B Duncan, et al. Hiv protease genotype and viral sensitivity to hiv protease inhibitors following saquinavir therapy. *Aids*, 12(13):1611–1618, 1998.

[222] Andrew R Zolopa, Robert W Shafer, Ann Warford, Jose G Montoya, Phillip Hsu, David Katzenstein, Thomas C Merigan, and Brad Efron. Hiv-1 genotypic resistance patterns predict response to saquinavir–ritonavir therapy in patients in whom previous protease inhibitor therapy had failed. *Annals of internal medicine*, 131(11):813–821, 1999.

[223] Dale J Kempf, Jeffrey D Isaacson, Martin S King, Scott C Brun, Yi Xu, Kathryn Real, Barry M Bernstein, Anthony J Japour, Eugene Sun, and Richard A Rode. Identification of genotypic changes in human immunodeficiency virus protease that correlate with reduced susceptibility to the protease inhibitor lopinavir among viral isolates from protease inhibitor-experienced patients. *Journal of Virology*, 75(16):7462–7469, 2001.

[224] H Van Marck, I Dierynck, G Kraus, S Hallenberger, T Pattery, G Muyldermans, L Geeraert, L Borozdina, R Bonesteel, C Aston, et al. The impact of individual human immunodeficiency virus type 1 protease mutations on drug susceptibility is highly influenced by complex interactions with the background protease sequence. *Journal of virology*, 83(18):9512–9520, 2009.

[225] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.

[226] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7): 990–999, 2016.

[227] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[228] Tanel Pärnamaa and Leopold Parts. Accurate classification of protein subcellular localization from high-throughput microscopy images using deep learning. *G3: Genes, Genomes, Genetics*, 7(5):1385–1392, 2017.

[229] Alessandro Ferrari, Stefano Lombardi, and Alberto Signoroni. Bacterial colony counting with convolutional neural networks in digital microbiology imaging. *Pattern Recognition*, 61:629–640, 2017.

[230] Andrea Arieno, Ariane Chan, and Stamatia V Destounis. A review of the role of augmented intelligence in breast imaging: from automated breast density assessment to risk stratification. *American Journal of Roentgenology*, 212(2):259–270, 2019.

[231] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

[232] Óscar Álvarez-Machancoses and Juan Luis Fernández-Martínez. Using artificial intelligence methods to speed up drug discovery. *Expert opinion on drug discovery*, 14(8): 769–777, 2019.

[233] Nic Fleming. How artificial intelligence is changing drug discovery. *Nature*, 557(7706): S55–S55, 2018.

[234] Simon I Hay, Dylan B George, Catherine L Moyes, and John S Brownstein. Big data opportunities for global infectious disease surveillance. *PLoS medicine*, 10(4):e1001413, 2013.

[235] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10, 2016.

[236] Amy Nelson, Daniel Herron, Geraint Rees, and Parashkev Nachev. Predicting scheduled hospital attendance with artificial intelligence. *NPJ digital medicine*, 2(1):1–7, 2019.

[237] Jonas C Ditz, Bernhard Reuter, and Nico Pfeifer. Convolutional motif kernel networks. *arXiv preprint arXiv:2111.02272*, 2021.

[238] Yanni Zhu, Xiaotong Shen, and Wei Pan. Network-based support vector machine for classification of microarray samples. *BMC bioinformatics*, 10(1):1–11, 2009.

[239] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002.

[240] Franck Rapaport, Andrei Zinovyev, Marie Dutreix, Emmanuel Barillot, and Jean-Philippe Vert. Classification of microarray data using gene networks. *BMC bioinformatics*, 8(1): 1–15, 2007.

[241] Cuilan Gao, Xin Dang, Yixin Chen, and Dawn Wilkins. Graph ranking for exploratory gene data analysis. In *BMC bioinformatics*, volume 10, pages 1–14. BioMed Central, 2009.

[242] Zheng Guo, Tianwen Zhang, Xia Li, Qi Wang, Jianzhen Xu, Hui Yu, Jing Zhu, Haiyun Wang, Chenguang Wang, Eric J Topol, et al. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC bioinformatics*, 6(1):1–12, 2005.

[243] Ian W Taylor, Rune Linding, David Warde-Farley, Yongmei Liu, Catia Pesquita, Daniel Faria, Shelley Bull, Tony Pawson, Quaid Morris, and Jeffrey L Wrana. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature biotechnology*, 27(2):199–204, 2009.

[244] Eunjung Lee, Han-Yu Chuang, Jong-Won Kim, Trey Ideker, and Doheon Lee. Inferring pathway activity toward precise disease classification. *PLoS computational biology*, 4(11): e1000217, 2008.

[245] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

[246] Li Chen, Jianhua Xuan, Rebecca B Riggins, Robert Clarke, and Yue Wang. Identifying cancer biomarkers by network-constrained support vector machines. *BMC systems biology*, 5(1):1–20, 2011.

[247] Sergei Vassilvitskii and David Arthur. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2006.

[248] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

[249] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997.

[250] Maxime Oquab, Léon Bottou, Ivan Laptev, Josef Sivic, et al. Weakly supervised object recognition with convolutional neural networks. In *Proc. of NIPS*, volume 2014, pages 1545–5963. Citeseer, 2014.

[251] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018.

[252] Guoqing Liu, Jianxin Wu, and Zhi-Hua Zhou. Key instance detection in multi-instance learning. In *Asian Conference on Machine Learning*, pages 253–268. PMLR, 2012.

[253] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[254] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

[255] Marcus Schmidt, Daniel Böhm, Christian von Törne, Eric Steiner, Alexander Puhl, Henryk Pilch, Hans-Anton Lehr, Jan G Hengstler, Heinz Kolbl, and Mathias Gehrmann. The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer research*, 68(13):5405–5413, 2008.

[256] Yudi Pawitan, Judith Bjöhle, Lukas Amler, Anna-Lena Borg, Suzanne Egyhazi, Per Hall, Xia Han, Lars Holmberg, Fei Huang, Sigrid Klaar, et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast cancer research*, 7(6):1–12, 2005.

[257] Yixin Wang, Jan GM Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679, 2005.

[258] Christos Sotiriou, Pratyaksha Wirapati, Sherene Loi, Adrian Harris, Steve Fox, Johanna Smeds, Hans Nordgren, Pierre Farmer, Viviane Praz, Benjamin Haibe-Kains, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, 98(4):262–272, 2006.

[259] Anna V Ivshina, Joshy George, Oleg Senko, Benjamin Mow, Thomas C Putti, Johanna Smeds, Thomas Lindahl, Yudi Pawitan, Per Hall, Hans Nordgren, et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer research*, 66(21):10292–10301, 2006.

[260] Christine Desmedt, Fanny Piette, Sherene Loi, Yixin Wang, Françoise Lallemand, Benjamin Haibe-Kains, Giuseppe Viale, Mauro Delorenzi, Yi Zhang, Mahasti Saghatchian d'Assignies, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clinical cancer research*, 13(11):3207–3214, 2007.

[261] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[262] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

[263] Yupeng Cun and Holger Fröhlich. Prognostic gene signatures for patient stratification in breast cancer-accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. *BMC bioinformatics*, 13(1):1–13, 2012.

[264] Elisabetta Pietri, Vincenza Conteduca, Daniele Andreis, Ilaria Massa, Elisabetta Melegari, Samanta Sarti, Lorenzo Cecconetto, Alessio Schirone, Sara Bravaccini, Patrizia Serra, et al. Androgen receptor signaling pathways as a target for breast cancer treatment. *Endocr Relat Cancer*, 23(10):R485–R498, 2016.

[265] Catriona Jamieson, Giovanni Martinelli, Cristina Papayannidis, and Jorge E Cortes. Hedgehog pathway inhibitors: A new therapeutic class for the treatment of acute myeloid leukemiahedgehog pathway inhibitors for acute myeloid leukemia. *Blood Cancer Discovery*, 1(2):134–145, 2020.

[266] Gillian Farnie and Robert B Clarke. Mammary stem cells and breast cancer—role of notch signalling. *Stem cell reviews*, 3(2):169–175, 2007.

[267] Jinhua Xu, Yinghua Chen, and Olufunmilayo I Olopade. Myc and breast cancer. *Genes & cancer*, 1(6):629–640, 2010.

[268] Xiang Li, Charles X Ling, and Huaimin Wang. The convergence behavior of naive bayes on large sparse datasets. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 11(1):1–24, 2016.

[269] Khanh Hiep Tran, Azin Ghazimatin, and Rishiraj Saha Roy. Counterfactual explanations for neural recommenders. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1627–1631, 2021.

[270] Shunsuke Kitada and Hitoshi Iyatomi. Attention meets perturbations: Robust and interpretable attention with adversarial training. *IEEE Access*, 9:92974–92985, 2021.